Lexical Levels and Formulaic Language:

An Exploration of Undergraduate Students' Vocabulary and Written Production of Delexical Multiword Units

by

Ruth Angela Scheepers

submitted in accordance with the requirements

for the degree of

Doctor of Literature and Philosophy

in the subject

Linguistics

at the

University of South Africa

Supervisor: Prof. E.J. Pretorius

Co-supervisor: Prof. E.H. Hubbard

November 2014

ABSTRACT

This study investigates undergraduate students' vocabulary size, and their use of formulaic language. Using the Vocabulary Levels Test (Laufer and Nation 1995), it measures the vocabulary size of native and non-native speakers of English and explores relationships between this and course of study, gender, age and home language, and their academic performance. A corpus linguistic approach is then applied to compare student writers' uses of three high-frequency verbs (*have*, *make* and *take*) relative to expert writers. Multiword units (MWUs) featuring these verbs are identified and analysed, focusing on delexical MWUs as one very specific aspect of depth of vocabulary knowledge. Student and expert use of these MWUs is compared. Grammatically and semantically deviant MWUs are also analysed. Finally, relationships between the size and depth of students' vocabulary knowledge, and between the latter and academic performance, are explored.

Findings reveal that Literature students had larger vocabularies than Law students, females knew more words than males, and older students knew more than younger ones. Importantly, results indicated a relationship between vocabulary size and academic performance. Literature students produced more correct MWUs and fewer errors than Law students. Correlations suggest that the smaller students' vocabulary, the poorer the depth of their vocabulary is likely to be. Although no robust relationship between vocabulary depth and academic performance emerged, there was evidence of an indirect link between academic performance and correct use of MWUs.

In bringing together traditional methods of measuring vocabulary size with an investigation of depth of vocabulary knowledge using corpus analysis methods, this study provides further evidence of the importance of vocabulary knowledge to academic performance. It contributes to debates on the value of a sound knowledge of high-frequency vocabulary and a developing knowledge of at least 5000 words to academic performance, and the analysis and quantification of errors in MWUs adds to our understanding of novice writers' difficulties with these combinations. The study also explores new ways of investigating relationships between size and depth of vocabulary knowledge, and between depth of vocabulary knowledge and academic performance.

Key words

Vocabulary size; Vocabulary depth; Vocabulary Levels Test; Corpus analysis; WordSmith Tools; High-frequency vocabulary; Delexical verbs; Multiword units

ACKNOWLEDGEMENTS

With thanks to my supervisors, Lilli and Hilton, for their erudition, their unfailing support, the countless hours they dedicated to helping me to finish this thesis, but most of all for their friendship;

to the Dean of the College of Human Sciences, for granting me financial assistance with which to collect the data;

to Alexa and Clea, for their generosity in helping me with the technical issues;

to Stella and Felicity and other friends and colleagues who helped me to survive the process;

but most of all to my daughters, Clea and Lucy, without whose love and encouragement I could not have done this.

DECLARATION

I declare that LEXICAL LEVELS AND FORMULAIC LANGUAGE: AN EXPLORATION OF UNDERGRADUATE STUDENTS' VOCABULARY AND WRITTEN PRODUCTION OF DELEXICAL MULTIWORD UNITS 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.

_____ _____

(Ms) R.A. Scheepers Date

Student Number: 06843905

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 INTRODUCTION

Knowing a language of course requires much more than simply knowing individual words and grammatical rules. It involves knowing how to use language in specific contexts, such as academic writing. Being able to write well in academic contexts requires breadth and depth of vocabulary knowledge (Hancioğlu, Neufeld and Eldridge 2008; Qian 1998; Read 2004, 2007), that is, the size of one's vocabulary, or the number of words one recognises, as well as the depth of this knowledge, which includes knowing which words typically keep company with which. This study developed from my feeling, growing over several years of teaching undergraduates in a university English department, that students, both native speakers (NSs) and non-native speakers of English (NNSs), have difficulty articulating their thoughts idiomatically in their academic writing. For example, in this study the corpus of writing by Literature and Law students features expressions such as the following:

he has **an aggressive behaviour** while he talk
he was **making a serious corruption**
some Africans also **were taken rescued** by the Spanish

Over and above signs of incomplete linguistic competence reflected in the errors of concord and tense and the confusion of articles, these examples highlight the difficulties such students have in producing appropriate lexical combinations, difficulties which are related to both the breadth and the depth of their vocabulary knowledge. For this reason, and as indicated in the title of the thesis, the study focuses on two specific aspects of vocabulary study, namely lexical levels and formulaic language. The term 'lexical levels' refers to the investigation of the size of students' vocabulary, measured using the Vocabulary Levels Test (VLT) (Laufer and Nation 1995), a test which quantifies active vocabulary knowledge according to five word-frequency levels. The term 'formulaic language', on the other hand, refers to the specific word combinations investigated in this study, the production of which, it is argued, reflects one very specific set of indicators of depth of vocabulary knowledge. While there are many different types of such word combinations referred to under the umbrella term 'formulaic language', the combination investigated in this thesis is a specific multiword unit (MWU) made up of a verb-noun collocation featuring a seemingly simple high-frequency verb, in this case *have*, *make* or *take*. In this study the size or breadth of students'

vocabulary knowledge and the depth of this knowledge, as reflected in their ability to produce MWUs featuring a combination of a high-frequency verb and a noun in their academic writing are investigated, using both a discrete-item vocabulary test and the tools of corpus linguistic analysis. Comparisons are also made between expert and student use of the three verbs in question, in course study material in the case of the former and in examination scripts in the case of the latter. Further to this, relationships between students' breadth and depth of vocabulary knowledge (as reflected in their use of MWUs) and their academic performance are explored.

This chapter provides the introduction to the thesis. The next section contextualises the study, and this is followed by a discussion of its main focal aspects, that is, size or breadth of vocabulary knowledge, and formulaic language in the form of MWUs. The nature of the study and the research questions are then discussed, followed by an introduction to the methodology adopted. The chapter concludes with a brief synopsis of the remaining chapters of the thesis.

1.2 CONTEXTUALISATION OF THE STUDY

This section provides some information on education in South Africa today in an attempt to contextualise the study.

In South Africa today, despite the fact that the democratic state is two decades old, low literacy and numeracy levels continue to hamper educational success (NEEDU 2013; Howie, Van Staden, Tshele, Dowse and Zimmerman 2012). This state of affairs, of course, affects students entering university. The National Benchmark Test (NBT) project (2009) which tested academic literacy and mathematics in about 13 000 students at several South African universities[1] revealed that, while almost 6000 students were 'proficient' (achieving between 64 and 100%) in academic literacy, almost as many were at an 'intermediate' (38–63%) level, and almost 1000 were at a 'basic' level (0–37%).[2] What these levels indicate is that students performing at the proficient level can be placed in a regular course of study at university and should succeed. Those scoring at the intermediate level may face challenges and will require extended programmes to succeed at university, while students scoring at the basic level are unlikely to succeed at university without 'extensive and long-term support' (NBT 2013). Thus, based on the NBT data (2013), more than half the students tested were unlikely to achieve success at university. Such students come predominantly from home, community and school contexts where there are limited literacy resources, few literacy practices or reading role models and low reading expectations. For instance, when interviewed between 2005 and 2009, 70% of primary school teachers in township schools had no more than 10 books in

[1] UKZN, Mangosuthu, Witwatersrand, Cape Town, Western Cape, Stellenbosch and Rhodes universities.
[2] http://www.nbt.ac.za/content/benchmark-levels [Accessed 21 December 2013].

their homes (Machet and Tiemensma 2009; Pretorius and Mampuru 2007:45; Pretorius and Ribbens 2005:146). This implies that there are low reading expectations in the classroom, and that teachers are not playing a strong role as reading models. By the time learners reach university, low literacy levels and poor schooling have already had a profound impact on their academic performance (Cooper 1999, 2000). They are unlikely to have read much independently and may be resistant to efforts to improve their reading.

This is something I have encountered in my own teaching: the students in this study are typical of first-year university students in South Africa. They were enrolled in semesterised first-year modules offered by the Department of English Studies at the University of South Africa (UNISA). This is an open distance learning institution, one that aims at

> bridging the time, geographical, economic, social, educational and communication distance between student and institution, student and academics, student and courseware and student and peers. Open distance learning focuses on removing barriers to access learning [sic], flexibility of learning provision, student-centredness, supporting students and constructing learning programmes with the expectation that students can succeed (UNISA Open Distance Learning Policy 2008).

These students come from a range of language backgrounds, including the nine official indigenous African languages, and for most of them English is a second or even a third language. Although many of them complete their high school education through the medium of English, the language in which instruction is conducted, their exposure to English is in many cases inadequate, with the result that by the time they reach university they have not mastered the written idiom and may have only a developing knowledge of academic English in particular. I have found that students are increasingly unwilling to read on their own, showing a reluctance to read even their prescribed books, even though the amount of text they are required to read in preparation for examinations has been pruned to the point where students in reality have to read very little in order to pass.

These students are products of a country in which there are in effect two systems of education. The first system comprises a fifth of all schools, the historically advantaged 'ex-Model C', Indian and private schools. These are the best resourced and best managed schools and in this system the majority of children (black and white) perform well on all literacy and numeracy measures. The second, much larger system is made up of 80% of schools. Here, the majority of learners perform poorly on all literacy and numeracy measures (Fleisch 2008). These are schools that have been historically disadvantaged (Reddy, Van der Berg, Janse van Rensburg and Taylor 2012; Spaull 2012; Taylor, Van der Berg, Reddy and Janse van Rensburg 2011). Today South Africa spends about 20% of its GDP on education in an attempt to restore the balance – far more than most middle-income countries and by far the most of African countries. Poor schools are now far better resourced than they were a few years ago. Yet literacy and numeracy levels remain low.

This has a deleterious effect on Grade 12 or matric outcomes. In the words of Taylor et al. (2011:4), '[t]he resounding verdict emanating from recent large-scale assessments of student achievement is that South African children are performing at worryingly low levels by international standards'. International assessments such as the Progress in International Reading Literacy Study (PIRLS) administered in 2006 (Howie 2010; Howie, Venter, Van Staden, Zimmerman, Long, Du Toit et al. 2008) found that South African children in Grade 5 were reading at lower levels than the Grade 4s in the 39 other countries included in the test (Taylor et al. 2011:4). In the application of PIRLS five years later in 2011, there was still no major difference in the overall achievement of South African learners (Howie et al. 2012:1). South Africa also performed very poorly in Grade 8 mathematics and science in the 2002 study of the Trends in International Mathematics and Science Study (TIMSS) (Taylor et al. 2011:4). Nine years later, TIMSS (2011) revealed little improvement: the tests were administered to Grade 9 pupils in South Africa, Botswana and Honduras, and in all three cases performance was low in both mathematics and science. South Africa's national scores were among the bottom six countries of the 42 countries tested at Grade 8 level, below the low-performance benchmark and, of these three developing countries, South Africa achieved the second lowest mean score in mathematics – Botswana 397, South Africa 352 and Honduras 338 – and the lowest score in science: Botswana 404, Honduras 369 and South Africa 332 (HSRC 2012:4).

South African students have not fared much better in assessments involving African countries only, such as the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) surveys. Surveys of Grade 6 reading and mathematics in 2000 and 2007 found that South African students were performing below average when compared to the other 13 Southern and Eastern African countries included in the surveys (Murimba 2005; SAQMEQ III 2010–2014; Taylor et al. 2011:7; Van der Berg 2008). The effect of these low levels of performance in literacy and numeracy at intermediate primary levels on matric outcomes was evident in a study which identified pupils who had been tested by TIMSS 2002 and who were in matric in 2006 or 2007 (Reddy et al. 2012:6–7; Taylor et al. 2011). It was found that, in the case of children from better off homes and attending historically advantaged schools, performance in mathematics in Grade 8 was strongly correlated with passing matric. The same could not be said for those children from lower socioeconomic areas and attending schools which had been historically disadvantaged. In fact, in 2007, while over 80% of white and Indian pupils between the ages of 21 and 25 had obtained their matriculation certificates, only 40% of black South Africans had done so. One in 11 white students achieved an A aggregate, while only one in 640 black students did so, and of this number almost half had attended historically advantaged schools (Taylor et al. 2011:3).

These circumstances are exacerbated by the apparent failure of many students to fully appreciate the extent of their language problems. Using data gathered in an earlier study from a question that asked students to rate their proficiency in English on a scale from 1 (poor) to 5 (successful, or good at), Coetzee-

Van Rooy (2011) found that while Afrikaans and African languages students in her study all overestimated their proficiency in English, the latter did this to a greater degree than the former. One of the reasons she provides for this is that while African students are in the main multilingual, and have skills in all the languages they are able to speak, it is often the case that the language they read in is English. They therefore tend to rate their proficiency in this language very positively because they compare their knowledge of this language and their ability to read in it to their ability to read in their home language, which is often less proficient as these languages are used more for social communication rather than educational purposes (Coetzee-Van Rooy 2011:169). Afrikaans students, on the other hand, are probably more likely to read frequently in Afrikaans as well as in English, with the result that their judgement of their English proficiency tends to be less inflated. Coetzee-Van Rooy found that this was supported by the data from her 2010 study reported on in this article (Coetzee-Van Rooy 2011). This inflated assessment of their English proficiency has implications for such students (Coetzee-Van Rooy 2011:171).

A second reason for this discrepancy between perceptions of English proficiency and test scores that the author (2011:171) provides is that when multilingual students rate their English proficiency they are really rating their 'ability to communicate in the language across cultures and languages' (2011:171). In other words, they are not assessing their ability to use the language to perform what she refers to as 'information tasks' (2011:171); for these students, there is little difference between spoken and written language or, to be more technical, between the distinction that Cummins (1999) makes between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP) (Cummins 1999).

The third reason for the discrepancy is that, compared to others in their communities, these students regard themselves as very proficient in English (Coetzee-Van Rooy 2011:171). Such students, like those taking part in my study, are often the first in their families to go to university and they have thus developed an inflated idea of their own proficiency that often translates into unrealistic estimations of their academic ability – because of the importance of English as the language of learning and teaching (LoLT) for African students. (2011:172). What Coetzee-Van Rooy (2011:173–4) also found was that, contrary to Dornyei's Motivational self System theory, where it is assumed that language learners are aware of the discrepancy between their 'actual L2 self', that is their actual proficiency, and their ideal L2 self, or their inflated idea of their proficiency, and that this awareness will translate into increased motivation and effort that will result in improvements in language learning, in Coetzee-Van Rooy's (2011) study this did not appear to result in motivation for higher achievement among the African students. Instead, she found that African students believed that they were already very proficient in English and it could thus be predicted that 'they would probably not be motivated to improve their skills in English language classrooms' (Coetzee-Van Rooy:174).

These findings have pedagogical implications for some of the students in the present study. While academics believe that proficiency in English contributes to academic success, students often do not believe that they have any difficulties or problems in this regard (Stephan, Welman and Jordaan 2004:42). The participants in Coetzee-Van Rooy's (2011) study indicated that they acquired additional languages in order to be able to communicate with others. They did not express the notion of learning the language for 'ideational purposes' (Coetzee-Van Rooy 2011:175). Thus they could argue that they were in fact highly proficient in English, but this was in effect an English used mainly for social communication. 'Their "desired" L2 self-image is in perfect harmony with the expectations of their peers, family and society in general' (Coetzee-Van Rooy 2011:175). This can go some way in explaining the fact that, despite the high status English has among African learners and their assumed motivation to learn the language, they achieve only low levels of proficiency. (Coetzee-Van Rooy 2011:175).

The situation thus becomes a vicious circle – such students enter university already at a disadvantage, and find it difficult, if not impossible, to catch up and close the gap, often falling by the wayside and becoming part of the 'drop out' statistics and affecting through-put figures.

Although this study is not primarily concerned with all these issues, they are the underlying reality of South African education and no study of this nature can fail to take them into account.

The following section provides a brief overview of the main focal points underpinning this study.

1.3 FOCUS OF THE STUDY

The main constructs underpinning this study are size, or breadth, and depth of vocabulary knowledge and formulaic language, in this case the phenomenon of multiword units (MWUs). These aspects are briefly dealt with in the following two sub-sections and will be discussed in greater detail in subsequent chapters.

1.3.1 Breadth and depth of vocabulary knowledge

This study investigates the relationship between the breadth of students' vocabulary knowledge, as measured by the VLT, and the depth of this knowledge as reflected in their ability to produce MWUs formed with the high-frequency verbs *have*, *make* and *take*[3] in their writing. Focusing on such high-frequency verbs may seem odd when dealing with students who should be advanced learners of English, and may also seem to overlook the importance of academic vocabulary to university studies. But breadth

[3] *HAVE* was chosen for investigation in this study because it was the most frequent high-frequency verb in the Student corpus; although *DO* occurred more frequently than *MAKE*, *TAKE*, *GET* and *GIVE* in the Student corpus, *MAKE* and *TAKE* were selected because they are known to be very productive of the MWUs in which I was interested (cf Biber et al. 1999) (see §5.3).

and depth of vocabulary, although measured using different techniques in this study, are inextricably linked, and as breadth of knowledge increases, so too should depth of knowledge (Milton 2013). Although such high-frequency verbs as these, and others such as *go, do, say, look, know, see, give, think, come, find, get* and *use,* are generally learnt early on in a learner's encounter with the language, the collocational behaviour of such words is not easy to learn or teach and may consequently be neglected by teachers, with the result that the deeper understanding of high-frequency words may not readily be achieved. However, such high-frequency vocabulary is vital to learners if they are to become independent readers. Research has found that a vocabulary of 3000 word families, or about 5000 words, is required for general reading comprehension and should allow readers coverage of 90 to 95% of the running words in any text (Laufer 1997). What this means is that depth of knowledge of very high-frequency words such as *have*, *make* and *take*, together with a breadth of knowledge of relatively high-frequency words (the 5000-word level), should provide learners with enough vocabulary to read authentic texts (Schmitt, Schmitt and Clapham 2001:56). Together with a developing knowledge of academic words, learners should be able to cope with university level reading and writing (this is discussed in more detail in Chapter 3).

However, as was discussed above, in South Africa many students enter university with low reading levels and a vocabulary that is not adequate for the tasks of academic reading and writing (§1.2). In her study, for example, Cooper (1999, 2000) found a relationship between the breadth of the vocabulary knowledge of first-year students at a South African university and their academic performance. Many of the students in her study lacked both the high-frequency and the academic word knowledge to cope with reading academic texts (Cooper 1999, 2000:28), and this would of course have repercussions for students' writing abilities. This situation is exacerbated by socioeconomic circumstances, poor schooling and print-poor environments (Machet and Tiemensma 2009; Pretorius and Mampuru 2007; Pretorius and Ribbens 2005). Add to this the complex nature of the lexicon and lexical issues with which students are faced and the fact that collocations such as *make a claim* and *take a decision* are common to academic writing and it becomes clear that this aspect of vocabulary knowledge is worth investigating further.

These lexical issues and the difficulties they pose for students are highlighted in the many different kinds of deviations which occurred in the writing of the students in this study, and in the differences between student writing and the writing by experts to which this was compared. Features such as an unawareness of concord (*most politician today have interest in government*), a simplified verb tense formation (*she establish him/herself and make contacts with the other*), unawareness of punctuation conventions (*at their second meeting they make love David liked her company*), and difficulties with lexis (*they come with this technology of making this notes*) were common. The deviations I was particularly interested in were those where students showed a lack of mastery of the restrictions placed on high-frequency verbs in collocations, deviations which rendered their language unnativelike and unidiomatic. Academic English requires a degree

of idiomaticity, a feature some of these students found difficult to achieve, often because of lexical or grammatical difficulties but sometimes through a lack of awareness of the 'patterns of a text' (Hunston and Francis 2000). Students are assumed to 'know' high-frequency words, but they often have difficulty using these correctly, or at least acceptably, in lexical phrases. They may choose unacceptable collocates, or may avoid these words altogether, using instead inappropriate or 'difficult' words, making their language stylistically incorrect, stilted or inappropriate. Few studies have explored the relationship between breadth and depth of vocabulary knowledge, specifically in relation to use of MWUs, and so my study is aimed at making a contribution to this important but under researched issue.

1.3.2 Multiword units

As a general term, MWUs refer to language units such as clusters of words, collocations and idioms, units of language that may be viewed along a continuum of idiomaticity, ranging from units completely free (*make dinner, take a book out of the library*) to those totally fixed in terms of syntax (*take care of*, *take advantage of*), and from units which are transparent (*have a look, make a decision*) to those which are obscure in terms of meaning (*curry favour*, *beat about the bush*) (Fan 2009; Howarth 1998a). The use of prefabricated MWUs in text is so pervasive that Sinclair (1991) proposed the 'idiom principle': 'a language user has available to him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments' (Sinclair 1991:110). This idiom-oriented view of language has been supported by the findings of studies on language processing (e.g. Pawley and Syder 1983) and corpus studies (e.g. Altenberg 1998; Sinclair 1991; Renouf and Sinclair 1991).

MWUs and the plethora of terms and definitions for the various combinations which fall under this umbrella term are described and discussed in some detail in Chapter 4. The combinations investigated in this study can be referred to loosely as collocations (e.g. Howarth 1998a, b). The importance of collocational knowledge in second language (L2) competence is beyond dispute (Fan 2009:111). Such knowledge allows learners to speak more fluently, makes their speech easier to comprehend and helps them to write and sound more native-like (Hunston and Francis 2000; Pawley and Syder 1983; Wray 2002). However, L2 learners' problems with collocational use have repeatedly been reported regardless of their level of language proficiency. For instance, Fan (1991, cited in Fan 2009) has reported on secondary students, Biskup (1992) on advanced learners and Farghal and Obiedat (1995) and Nesselhauf (2003, 2005) on university students. One of the difficulties students encounter with these combinations lies in the idiosyncratic nature of collocational use (e.g. *strong* has a similar meaning to *powerful* in most instances but *strong* collocates with *tea* but not with *car*). Another difficulty arises from the fact that collocational use may be very different across languages.

In Hunston and Francis's (2000:1) explanation of the phenomenon of collocation, a word comes with its attendant phraseology – the grammar pattern that belongs to that particular word. According to these researchers, all words can be described in terms of these patterns. Hunston and Francis (2000) believe that our experience of a language allows us to understand these patterns in a text. But this intuition is not always reliable – and in the case of many of the students in my study, their experience of the language was often limited, making it consequently inadequate in providing them with this awareness. Hunston and Francis (2000:3) believe that patterns and lexis are 'mutually dependent in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns' (2000:3). These patterns have an important function in creating meaning in that the meaning of a word is often distinguished by its typical occurrence in various patterns, and words which share a particular pattern also often share an aspect of meaning.

In this study, the focus falls on MWUs formed with three high-frequency verbs, *have*, *make* and *take,* and how their use in student writing compares to that in expert writing, in Literature and Law courses. These three verbs were chosen for this study because they are highly polysemous, that is, they have two or more closely related meanings, caused by the fact that they can be used generally in more abstract, delexicalised or grammatical uses, but that they also have various language-specific tendencies resulting in specialised meanings, collocations and idiomatic uses (Viberg 1996, cited in Altenberg and Granger 2001). All languages have such basic verbs which are used very often and which regularly top frequency lists (Altenberg and Granger 2001). Research has shown that these high-frequency verbs tend to be problematic for foreign or L2 learners (Altenberg and Granger 2001: 174; Kaszubski 2000; Lee and Chen 2009:155; Wang and Shaw 2008; Yan 2006). They are highly productive of the type of combination under study here, combinations such as *make a decision*, *have a look* and *take a journey.*

Verbs such as *have, go, do, say, take, give, get* and *make* have been variously referred to as *light verbs* (Live 1973; Wittenberg and Piñango 2011), *small verbs*, *support verbs* (Langer 2004) and *delexical verbs* (Altenberg and Granger 2001; Howarth 1998a, b). This is because of their tendency to be found in combinations where the noun carries the main semantic weight of the expression. Howarth (1998a) explains that these high-frequency verbs are often used delexically, that is, in such a way that their meaning is defined by the company they keep (an expression coined by Firth (1957, cited in Léon 2007; Howarth 1998a), while they themselves carry little meaning. The words they keep company with (with which they collocate, or the words that make up the lexical chunks or MWUs) are often not arbitrary, but in fact what Howarth (1998a, b) refers to as 'restricted' collocations. These are the combinations referred to in this study as *delexical multiword units*, which Algeo (1995) calls *expanded predicates,* which Nesselhauf (2005) calls *stretched verb constructions*, and which Langer (2004) calls *support verb constructions*.

Nattinger and DeCarrico (1992) refer to a continuum, as does Howarth (1998a, b) in the amount of variation that occurs in lexical phrase patterns. They distinguish between lexical phrases, which they regard as collocations with pragmatic functions (*how do you do?*), and collocations, which are what Howarth (1998a, b) would call idioms, with no pragmatic function (Nattinger and DeCarrico 1992:36). Nattinger and DeCarrico suggest that the less variation a combination of words allows, the more predictable and easier it is, and the easier it is to acquire. This accounts for much of the way language is processed: 'the degree to which words constrain those around them, and the assurance we have that certain words are going to follow certain others, are the facts we use to create all sorts of subtle variations and surprises' (Nattinger and DeCarrico 1992:34).

In this study, the data comprised the course material used in the UNISA Literature and the Law courses, making up the Expert corpus, and the essays written in the end-of-semester examination by the students in the sample, making up the Student corpus. The MWU in question is regarded as a type of collocation in that it is made up of a chunk of language, somewhere on the continuum between free combinations and idioms; that is, it falls into Howarth's (1998a, b) category of restricted (or semi-restricted) collocations in that some substitution on choice of the verb or of the noun is allowed in most cases. These MWUs are made up of

- a high-frequency verb, often used delexically, conveying little meaning itself, and
- an eventive noun which carries the semantic weight of the expression, and may often be a lower frequency, academic word.

These combinations include examples such as *have a look*, *make a comment*, *take a turn*. They are classified as core or pseudo delexical MWUs, discussed in more detail in Chapter 4.

The sections above have provided some context which is important to understanding the educational issues that have informed the study, and information on aspects which are central to the study. As noted, the study investigates both the breadth and the depth of students' vocabulary knowledge; it compares the size of students' vocabulary according to several variables: course of study, gender, age and language background. It also explores a particular aspect of vocabulary depth, namely how students' depth of vocabulary knowledge is expressed in the use of selected MWUs and how this use compares between students and expert writers in Literature and Law courses. In the following sections of this chapter, the rationale for the study, the research aims and research questions and the methodology used to address them are briefly discussed.

1.4 RATIONALE FOR THE STUDY

It is widely accepted that MWUs play a major role in how we process and use language (Nesselhauf 2005:1). The fact that such MWUs cause learners difficulties is also well attested (Howarth 1998a, b;

Nesselhauf 2003, 2005; Wang and Shaw 2008; Yan 2006). This study seeks to establish whether this is also reflected in the writing of the students who made up the sample in this study, and to explore the relationship between the vocabulary these students know and aspects of their production of these word combinations. To this end, the study compares the writing of a specific group of South African university students with that of expert writers in the same fields. The difficulties of such students require further investigation if aspects of language use which pose particular challenges to them are to be identified. As appropriate and idiomatic use of language involves knowledge of words and the company they keep (Firth 1957, in Léon 2007), it is reasonable to assume that the more vocabulary learners have acquired, the more collocations they are likely to use appropriately. Studies have shown that many collocations used in academic English writing feature high-frequency words (Algeo 1995; Altenberg and Granger 2001; Howarth 1996, 1998a, b; Langer 2004; Lennon 1996). Thus, in order to achieve competence in L2 collocational use, learners need to develop not only a wide vocabulary and sound knowledge of collocations (Fan 2009) but also an understanding of the many uses of the most basic high-frequency verbs. The MWUs investigated in this study are made up of a high-frequency verb and a noun, and the fact that the latter is frequently an academic word is a further indication of the relevance of both breadth and depth of vocabulary to success in university study. 'Academic words' are words that are commonly used in academic discourse across discipline domains, e.g. *hypothesis, category, postulate*. These are the sort of words captured in the University Word List (UWL) (Xue and Nation 1984) and in the Academic Word List (AWL) (Coxhead 2000). Studies have highlighted the need for students to master academic discourse in order to succeed in tertiary studies (Flowerdew 2001:371; Granger and Rayson 1998); this includes a thorough knowledge of such academic vocabulary.

Observations such as these have prompted the present study, which applies the methods of discrete-item vocabulary testing as well as those of corpus analysis in an attempt to arrive at a description of a group of students' vocabulary knowledge and how this relates to their production of MWUs containing a high-frequency verb used delexically. The study was informed by the theoretical frameworks of both corpus-based linguistics and corpus-driven linguistics, arriving at a method which is increasingly recognised in corpus studies as being a kind of hybrid (Biber 2009). As such, the analysis takes a corpus-driven approach, that is, what McEnery and Hardie (2012:147) refer to as 'corpus-as-theory' or the neo-Firthian approach, using an untagged corpus, but at the same time borrows elements from corpus-based methods, that is, 'corpus-as-method' (McEnery and Hardie 2012:147), or an approach in which the corpus is used 'mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study' (Tognini-Bonelli 2001:65–66). This hybrid approach is used increasingly in present-day corpus linguistic research and is thus well founded in the discipline.

This study does more than simply look at differences between student and expert writing; it is also an attempt to make a contribution to both vocabulary studies and corpus linguistics by exploring the relationships between breadth of vocabulary knowledge, an aspect of the depth of this knowledge and academic performance in a group of typical South African undergraduate students.

1.5 RESEARCH AIMS AND RESEARCH QUESTIONS

As stated, this study explores the relationship between the size of students' vocabulary (a vital component of their linguistic competence) and an aspect of the depth of their awareness of certain phraseological norms of academic English (Howarth 1998a, b), to the extent that this is reflected in their production of specific delexical MWUs. In this study the term 'size', as the more common term in the literature when referring to measurement and actual vocabulary tests, is used; 'breadth', as a rather more abstract concept, like 'depth', is used mainly where the distinction between breadth and depth of vocabulary knowledge is made. The first aim of the study was to measure the breadth and depth of students' vocabulary knowledge and to explore the relationship between this knowledge and their academic performance. A second aim was to investigate the relationship between the variables of course of study, gender, age and language background, and students' vocabulary size. A third aim was to establish the difference in depth of vocabulary knowledge between student and expert writers in two courses, Literature and Law, to the extent that it may be measured by the production of specific MWUs. The final aim was to explore the relationship between students' production of these specific MWUs and their academic performance. These aims were addressed by answering six main research questions (see §5.2 for a more detailed discussion). In doing so, the study fell naturally into three phases.

Phase 1 of the study begins with a consideration of discrete items of vocabulary (the more traditional way of investigating the vocabulary of learners) and the establishment of the relationship between measures of size of students' vocabulary and variables such as course of study, gender, age and language background, and between measures of breadth and their academic achievement. The research questions which gave rise to this process were expressed as follows:

Research Question 1

What is the size of the productive[4] vocabulary of undergraduate students?

[4] This study measures productive vocabulary, that is students' ability to use words in writing, also called active vocabulary knowledge. This differs from receptive or passive vocabulary knowledge, which is the ability to recognise words when reading.

This question was broken down into four sub-questions in which the relationship between vocabulary size and the variables of course of study (Literature or Law), gender, language background and age were measured (see §5.2).

Research Question 2

The second main research question in this phase explored the relationship between breadth of vocabulary knowledge and academic performance and was expressed thus:

What is the relationship between the size of students' productive vocabulary and their academic performance?

The focus then moves to Phase 2 and the use of the three verbs, *have*, *make* and *take* in both Student and Expert corpora. The two main research questions addressed in this regard were expressed as follows:

Research Question 3

How does the distribution of the functions of the three selected verbs compare within and across the Expert and Student corpora?

This question was broken down into four sub-questions (see §5.2).

This leads on to a further investigation of aspects of students' depth of vocabulary knowledge through the investigation and analysis of corpus evidence of their collocational knowledge, as expressed in their use of MWUs:

Research Question 4

How does students' use (in terms of frequency and deviance) of selected MWUs compare with the use of these MWUs by expert writers, within and across courses?

This question was broken down into three sub-questions in which comparisons of the use of MWUs in the Expert sub-corpora and in the Student sub-corpora, and between the two Student sub-corpora were made (see §5.2).

In the final phase, Phase 3, the link between size of vocabulary and the use of delexical MWUs is investigated, as well as the relationship between these two measures of vocabulary knowledge and academic performance (measured by examination scores). The two main research questions driving this last phase were expressed as follows:

Research Question 5

What is the relationship between the size of students' productive vocabulary and their production of selected MWUs, within and across courses?

Research Question 6

What is the relationship between students' production of selected MWUs and their academic performance, within and across courses?

This study thus brings together the two aspects of vocabulary study, breadth and depth of vocabulary knowledge, and their relationship in terms of academic performance, something which is as yet relatively rare in the research literature.

1.6 METHODOLOGY

In this section, the methodology followed in this study is briefly discussed. It is covered in greater detail in Chapter 5.

The sample from which data were collected was made up of undergraduate students enrolled at a South African 'open distance' university. This university is the largest distance education institution in the southern hemisphere and, as a result, the student population is heterogeneous: students come from all walks of life and have diverse school backgrounds. As it is an open distance learning institution, students vary greatly in age, linguistic and socioeconomic background and place of residence. All students study at a distance, via correspondence and electronically, with some living beyond the borders of South Africa (see §1.2). This makes for a very varied student population. The study population comprised the entire body of students registered for three modules in the Department of English Studies in the first semester of 2010, some 4500 in total. These three courses were the two modules making up the first level of English Literature studies (ENN101D and ENN102E), and a service course for students studying law, English Communication for Law Students (ENN106J). The two literature courses were combined so that there were two groups, Literature students and Law students. These courses were chosen because they presented examples of different study material and different writing by students – critical essays about literature in the case of the literature course, and argumentative essays about legal concepts in the case of the law course. The students in these courses were also representative of typical undergraduates and, finally, the courses were familiar and accessible to the researcher.

Phase 1 is quantitative in design as it is concerned with measuring the size of students' productive vocabulary and examining its relationship to various factors such as course, gender, age and language

background, and to examination scores. Students' breadth of vocabulary knowledge was measured using Version 2 of Laufer and Nation's (1995) Vocabulary Levels Test (VLT – the Active version) (Schmitt et al. 2001) in order to establish the levels of vocabulary at which students had achieved mastery, and those levels where their knowledge was still developing. A questionnaire elicited information on students' education and language background; further biographical details, namely age, gender and ethnicity[5], were extracted from university records. The software SPSS (IBM SPSS Statistics Version 21) was the main source for the computations and statistical analyses in this study. In order to establish whether there were relationships between students' scores on the VLT, both in its entirety and at the various levels, and their examination scores, correlations and regression analyses were performed. ANOVAs were used to examine differences between groups in terms of course, gender, age and language background.

This study also falls within the discipline of corpus linguistics in that it is concerned in part with the analysis of two corpora of academic English (discussed briefly here, and in greater detail in Chapters 2 and 5). In the past, most vocabulary studies involved the testing of knowledge of discrete vocabulary items – in a previous study, for instance, I investigated the knowledge that learners in an immersion situation had of individual words (Scheepers 2003, 2006). However, De Beaugrande (2000, cited in De Beaugrande 2001:11), for example, believes that data from corpora are able to reveal how a language is made up not of '*a closed set of formal rules*' determining '*well-formed strings*'. Instead, a language is made up of an '*open repertory of rich guidelines* for constructing *meaningful discourse events*'; the 'most essential trait' of language is to be '*selectable* and *combinable*', rather than '*well-defined* or *distinctive*' (De Beaugrande 2000, cited in De Beaugrande 2001:11). Corpus research has identified two ways in which language is selectable and combinable, namely through grammatical combinations or colligation, and through lexical combinations called collocations. The colligability and collocability of combinations of words influences whether discourse sounds 'fluent, natural or idiomatic' (De Beaugrande 2001:12) to the native speaker, and it is these aspects of selection and combination that non-native speakers frequently find difficult to master.

Thus Phase 2 of the study investigates language as it occurs in authentic discourse, gathered from the writing of both native speakers and non-native speakers from various language backgrounds. The explorative, qualitative methods of corpus linguistic analysis, in the sense of manual exploration and interpretation of concordance lines, as well as quantitative techniques, were used to examine the distribution of functions of the three verbs within and across the Expert Literature and Expert Law corpora and the Student Literature and Student Law corpora (these names are capitalised in order to indicate the specific corpora and genres). Once the process of manual exploration and quantification had been completed, the use of delexical MWUs in these corpora was compared.

[5] Information on ethnicity or race group is required by government. Such information is used in matters of redress of previously disadvantaged groups, for instance, and was included here to help to explain disparities in performance by different groups of students.

In order to collect the data that would make up these corpora, I sent a tutorial letter to all students enrolled in the Literature and Law first-year courses I had targeted, that is, ENN101D and ENN102E (these two modules together formed English First Level, and will henceforth be conflated into 'Literature'), and ENN106J, an English communication service course catering for students studying law, referred to as 'Law' in the study. The tutorial letter included a copy of Laufer and Nation's (1995) VLT, a questionnaire and a letter explaining the study and requesting students' participation. Ethical concerns (see §5.11) were addressed in this letter, as was the questionnaire. Students were provided with a self-addressed envelope with postage paid and asked to send the signed letter, the completed questionnaire and the completed VLT back to me by a specific date. Thus there was no invigilation during the completion of the VLT; students were relied upon to answer the test as best they could, without recourse to a dictionary or any other help (see Chapter 5). I was aware that this presented a potential limitation of the study, but there was unfortunately no alternative to this method of data collection in the given research context.

At the end of the semester, the examination scripts of those students who had returned the test and questionnaire were drawn, typed and compiled into the Student corpus. This was thus a collection of writing by undergraduate students of English at a South African open distance learning (ODL) university, comprising essays written under timed examination conditions. This corpus includes the writing of L1 English speakers as well as the writing of speakers of an indigenous African language or of Afrikaans who speak English as a second language.

The second corpus used in this study is the Expert corpus. This comprised the study material written by experts in these disciplines (i.e. Literature and Law), and which the students read in their English studies.

Once these corpora had been compiled, they were analysed and compared using the software program WordSmithTools (Version 5) (Scott 2008), described in detail in Chapter 5. Among other features, it allows one to establish word frequencies, to generate concordances and to search for keywords. In this way patterns in the use of the selected verbs and the MWUs were identified. The findings are reported on in detail in Chapter 6 (§6.3).

In the last part of the study, Phase 3, relationships between breadth of vocabulary knowledge (as revealed in students' scores on the different levels of the VLT and their total scores on the same test) and students' production of MWUs in their writing (a measure of the depth of their vocabulary knowledge), and their academic performance (measured by their scores in the end-of-semester examination) were investigated. This part of the analysis of the data was, like the first phase, quantitative. SPSS was used to examine relationships between vocabulary size and the production of MWUs. This was established by extracting texts from a sample of students in Literature and in Law whose scores were closest to the 25th, 50th and 75th

percentiles of the examination scores. Correlations were performed in order to test for significant relationships between breadth of vocabulary knowledge and depth of vocabulary knowledge expressed in students' production of MWUs, and between depth of vocabulary knowledge and examination scores (an indicator of academic proficiency).

1.7 STRUCTURE OF THE THESIS

This chapter has provided some background to the study and given a brief explanation of its focus and its rationale. Finally, it has set out the research aims and questions and outlined the methodological approach.

The literature review spans three chapters. Chapter 2 is concerned with the discipline of corpus linguistics and the analysis of computerised corpora. Through a review of relevant research in this area the chapter demonstrates how I reached my understanding of the approach followed in this study, that is, through a discussion of the two approaches developed by earlier researchers, the corpus-based (corpus-as-method) idea, and the corpus-driven (corpus-as-theory) idea espoused by the neo-Firthians, to what Biber (2009) called a hybrid approach, which takes elements from both the corpus-based and the corpus-driven schools. Although corpus linguistics relates most specifically to Phase 2, this aspect is discussed first as it also informs other aspects of the study.

Chapter 3 reviews vocabulary studies and relates particularly to Phase 1. It looks at how studies have moved from those taking a discrete-item perspective, to studies in the 21st century which consider vocabulary in the context of bodies of text, that is, corpora. It also reviews what research has revealed about the type of words that are most useful to learners at university level.

Chapter 4 discusses the aspect of vocabulary with which Phase 2 of this study is concerned, multiword units (MWUs). There has been a great deal of research into this aspect of vocabulary in recent years and the chapter attempts to make sense of the multitude of definitions and explanations that researchers have provided. This chapter relates both to Phase 2, where multiword combinations in the corpora are investigated, as well as to Phase 3, where the relationship between breadth of knowledge of vocabulary and the production of MWUs is investigated.

In Chapter 5 the methodology followed in this study is discussed in detail. The pilot study is described and the three phases of the main study and the methods used in each are explained, including quantitative statistical analyses in Phases 1 and 3, where SPSS is used to analyse the data, and both quantitative and qualitative corpus linguistic analyses in Phase 2. In this phase, WordSmith Tools version 5 (Scott 2008) is used to investigate the data and extract the MWUs for analysis.

Chapter 6 presents the results of the analyses and the discussion of these findings. Again, the chapter follows the pattern of the entire thesis, presenting findings in line with the three phases of the study.

Chapter 7 concludes the thesis. It provides a brief review of the aims and content of the study, and summarises the findings. The contributions made by the study to the field are discussed, the implications of the findings are considered and recommendations are suggested, based on these findings.

Chapter 2

Corpora and Corpus Linguistics

> Corpus data now signal some 'missing links' between these two sides [language and discourse] through regularities that are more specific than language and more general than discourse. These regularities could help to explain why native speakers of English sound 'fluent' and why their usage sounds 'natural' or 'idiomatic' (De Beaugrande 2001:3).

2.1 INTRODUCTION

This study aims to examine relationships between the vocabulary size of a sample of university students, their use of multiword units (MWUs) of a particular type in their academic writing and their academic performance. In order to do this, corpora of both student and expert writing were collected and analysed and, as such, the study falls within the discipline of corpus linguistics.

This chapter discusses the development of corpus linguistics and positions this study within this particular area of linguistic studies. It starts by explaining the concept of 'corpus', and developments that have taken place in the creation and compilation of corpora. It goes on to provide a brief history of the development of the discipline and discusses some of the individuals who have played a major role in this development. It then moves on to a discussion of some of the current debates in corpus linguistics. Finally, it discusses some corpus studies that are relevant to the present study, both from South Africa and elsewhere.

2.2 CORPUS LINGUISTICS

Before one can discuss corpus linguistics in any detail, it is necessary to define its most important element, the corpus itself.

2.2.1 What is a corpus?

There are many and diverse views on what constitutes a corpus, as well as many variations in the definitions provided by scholars. Although Leech (1992, cited in Tognini-Bonelli 2001:53) refers to a corpus as 'a helluva lot of text, stored on a computer', a corpus is not simply an arbitrary collection of texts, as this 'definition' may seem to imply. In fact, there are several important issues that corpus builders should keep

in mind (see Chapter 5), although these are still under discussion in some cases. In her study, Tognini-Bonelli (2001:55) regards a corpus as

> a computerised collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language.

In their definition of corpus linguistics, McEnery and Hardie (2012:1–3) observe that it is research which uses a collection of texts in electronic format, which is therefore machine readable (that is, by a computer), although there are exceptions as they note, where the data may be in written form or even in video form. In this context, 'texts' refers to a file of data which can be read by computer; data that are grouped together but retain their individual identity. Such a collection of texts is usually too large to be analysed manually and this has led over the years to the development of tools with which to investigate the data. The most common type of tool is the concordancer, such as WordSmith Tools (Scott 2008) which, like many other researchers, I have used in this study. Concordancers allow researchers to study words in context, and also to extract frequency data. These two aspects comprise the two types of analysis that are vital in corpus linguistics, that is, qualitative and quantitative (McEnery and Hardie 2012:1–2). Most concordancers allow researchers to find words in the context of the corpus as a whole, but also to pinpoint the source of a particular stretch of language (McEnery and Hardie 2012:1–2). This ability to identify the source text of particular words or combinations of words in the corpus is an important capability and one which I made use of in Phase 3 of the study, where comparisons are made between individual students' scores on the VLT and their use of MWUs.

2.2.2 The 'generations' of corpora

Corpora first appeared in the first three decades of the twentieth century, long before Sinclair's seminal work or present-day studies such as Tognini-Bonelli's. During the so-called 'first generation', compilation meant the manual extraction of texts from books and other printed matter, and a corpus of a million words in the 1970s was considered to be very large indeed. Kennedy (1998:17) refers to Jesperson, Kruisinga and Poutsma, all of whom used corpora as a basis for their grammars in the first quarter of the previous century. With the advent of computers and general access to these, data could be typed and saved on a computer and much larger corpora could be compiled. By now (the second decade of the 21st century) we are in what Moon (1997) has called the 'third generation' of corpora; the development of technology to scan texts electronically has meant that immense numbers of words can be collected, and this means too that these corpora can be regarded as reasonably accurate representations of the English language (Schmitt 2000:68–69). Such huge or 'mega' corpora (Kennedy 1998) may contain hundreds of millions of words, such as the COBUILD Bank of English Corpus (Sinclair 1987), the Cambridge International Corpus

(CIC)[6] and the British National Corpus (BNC)[7], to name only a few (these and other corpora are discussed in more detail below).

2.2.2.1 Early corpora

Two examples of early corpora are the Brown University Corpus (Francis and Kučera 1964) of American English, and its British English equivalent, the Lancaster-Oslo/Bergen Corpus (LOB) (Hofland and Johansson 1982). In 1963, Nelson Francis, Henry Kučera and others at Brown University compiled a sample of present-day American English for a computerised corpus. This was originally conceived as a 'snapshot' corpus (McEnery and Hardie 2012:97), using American English from the year 1961. This later became known as the Brown Corpus. It was made up of 500 samples, each of about 2000 words of continuous written English, and contained in total about 1 014 300 words, with samples taken from a range of text categories (Kennedy 1998:24). The sampling framework used to compile this corpus has become the 'de facto standard' when compiling small-scale written corpora (McEnery and Hardie 2012:97). As the first electronic corpus to be made widely available, the Brown Corpus set a precedent for free access and it has been available at no cost to researchers ever since its compilation, without any copyright fees.

The Lancaster Oslo/Bergen (LOB) Corpus (Hofland and Johansson 1982; Schmitt 2000:69) emerged some 10 years later and was intended as a British counterpart to the Brown Corpus, sampling as it did British English from 1961. This meant that reliable comparative studies of synchronic variation were now possible using these two 'matching' corpora (Kennedy 1998:28–29).

Other corpora were subsequently compiled and modelled on the Brown Corpus. These are known collectively as the 'Brown Family' and include samples of other varieties of world Englishes, samples from other later, and some earlier, years than 1961. This group of 'matching' corpora (McEnery and Hardie 2012:98) has made diachronic studies possible. The core corpora in this group, such as the Brown, LOB, Frown and FLOB – the latter two developed at the University of Freiburg by Christian Mair and his team (Hundt et al. 1998, 1999, cited in McEnery and Hardie 2012:99) – allow for simultaneous synchronic and diachronic comparisons, whereas the more marginal corpora in the group allow only for either synchronic or diachronic comparisons (McEnery and Hardie 2012:98). The latter corpora include the Kolhapur Corpus of Indian English (Shastri, Patilkulkarni and Shastri 1986), the Wellington Corpus of English in New Zealand (Bauer 1993) and the Australian Corpus of English (ACE) (Collins and Peters 1988), also known as the Macquarie Corpus of Written Australian English. These corpora contain only written English.

[6] http://www.cambridge.org.br/
[7] http://www.natcorp.ox.ac.uk/

2.2.2.2 Second generation corpora

Second generation corpora developed as technology advanced and it became possible to capture and store much larger collections of texts. By the end of the twentieth century huge corpora of 100 million words or more had become the norm (Kennedy 1998). During the early to mid-seventies, the London-Lund Corpus (LLC) was produced at Lund University in Sweden under the leadership of Jan Svartvik, containing a hundred 5000-word texts, about half a million words, and this was the largest collection of electronically available spoken English until well into the nineties (Kennedy 1998:31–32). This corpus was made up of the spoken part of Quirk's Survey of English Usage (SEU) which was started in 1959 and which recorded both written and spoken English, sampled according to genres and contexts (McEnery and Hardie 2012:74).

By far the most important of the second-generation corpora were the COBUILD Project, the Cambridge International Corpus (CIC) and the British National Corpus (BNC). The COBUILD Project was the first major mega-corpus project and was a joint venture between the publishers Collins and a research team from the English Department of the University of Birmingham, whence its name, COBUILD (Collins-Birmingham University International Lexical Database). It provided data for the development of a new English dictionary and began in 1980 under the leadership of John Sinclair (McEnery and Hardie 2012:80). Unlike its predecessors such as the Brown Corpus (Francis and Kučera 1964) and the LOB and the Lund Corpora, which used sampling methods to select a sample of texts, Sinclair's COBUILD Corpus included only whole texts.

2.2.2.3 Mega or third-generation corpora

In 1991, owing to the success of the COBUILD project, the development of a huge monitor corpus began, known today as the Bank of English corpus. A monitor corpus is 'designed to track current changes in language' and changes rapidly in size as texts are added to it annually, monthly and even daily (Hunston 2002:16). By 1997 this corpus was reported to contain over 300 million words, and to be still growing (Kennedy 1998:47; McEnery and Hardie 2012:80). This belongs to what Moon (1997) has referred to as the 'third generation' of corpora.

Today, there are many huge electronic text databases containing hundreds of millions of words available to researchers, besides the ones already mentioned – in fact, too many to estimate a number at this point. These are electronic collections of machine-readable text, usually unstructured in that they have not been compiled to represent specific categories of text, or annotated in any way, but available, often freely, for researchers to use should they so wish. Examples are the seven corpora created by Mark Davies, professor of Linguistics at Brigham Young University, and used by more than 100 000 people a month. These corpora include the *Corpus of Contemporary American English* (*COCA*), comprising 450 million words, the *Corpus of Historical American English* (*COHA*), 400 million words, and the new *Corpus of American Soap Operas* with

100 million words. Also available online are the British Academic Spoken English Corpus (BASE), developed at the Universities of Warwick and Reading, under the directorship of Hilary Nesi, and made up of over one and a half million words from 160 lectures and 39 seminars in various disciplines, and available from the Warwick Centre for Applied Linguistics.[8] The British Academic Written English (BAWE) corpus was created through collaboration between the Universities of Warwick, Reading and Oxford Brookes, funded by the Economic and Social Research Council. It is made up of 6.7 million words in texts written by British students.[9] As Kennedy (1998) observes, it is impossible to predict how these huge databases may be used in thematic, stylistic, lexicographical, historical or other linguistic studies in the future.

2.3 CORPORA FOR SPECIFIC PURPOSES

Many different types of corpora have been developed over more recent years as corpus linguistics has become more and more widespread. The sections below discuss corpora designed for specific purposes; these include corpora of varieties of English and learner corpora. The latter are not always concerned merely with L1 versus L2 learners; as in the present study, learners may be categorised uniquely according to the individual study. In my study, 'learners' is used to refer to the students in the sense that they are novice writers learning to write in the academic idiom. In fact, they come from a variety of language backgrounds, as discussed in Chapter 1 (§1.2). Flowerdew (2001:363) observes that, since the beginning of the 1990s, corpus linguistics has expanded in two directions: there are now more and much bigger (mega) corpora (§2.2.2.3), but at the same time there are also many more smaller corpora, often genre-based and designed for specific purposes, many initiated by Biber (1988, cited in Flowerdew 2001:364). These smaller corpora generally consist of between 20 000 and 200 000 words and tend to be more specific than larger corpora as far as topics or genre are concerned (Aston 1997), and are mostly scientific or technological in nature and feature written academic discourse.

It is increasingly common today for researchers to make use of smaller corpora that can be stored and analysed on a personal computer and still provide reliable and useful information about language. Kennedy (1998) believes that there is certainly a place in the discipline for corpora of fewer than one million words, and that these can provide useful insights, particularly for the teaching and learning of language. This provides some justification for the use of smaller corpora in the present study, as will be discussed in more detail in Chapter 5 (§5.10).

However, the majority of these smaller, specific corpora are made up of native speaker or 'expert' writing – not many learner corpora are as yet available commercially, and teachers and researchers have been

[8] http//www.reading.ac.uk/AcaDepts/ll/base_corpus/index.htm
[9] http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/

obliged to collect their own. However, Flowerdew (2001:364) and others (Guillot 2005; Nesselhauf 2005; Shirato and Stapleton 2007) feel strongly that the only way to gain insight into learner difficulties and to design materials that really address their needs is to use learner corpora in addition to expert ones. Learner corpora are discussed in more detail in §2.3.2 below.

2.3.1 Corpora of varieties of English

Today there are many corpora of varieties of English, spoken and written by speakers with first languages other than English. The International Corpus of English (ICE) is the largest corpus currently available for the comparative study of varieties of English (McEnery and Hardie 2012:74) and was started by a team led by Sidney Greenbaum, Director of the Survey of English Usage (SEU, see §2.2.2.2) at University College London (UCL) in 1988. At its inception it was envisaged that there would be 20 parallel corpora, each of a million words of English used by adults over the age of 18 who had received formal education through the medium of English. Today this corpus represents varieties of English which Kachru (1986, cited in McEnery and Hardie 2012:100) referred to as 'inner circle', that is, English spoken as a first language, 'outer circle', or English spoken as a second language, and 'expanding circle' English, that is, English spoken as a foreign language (Kennedy 1998:54–55). To date, while there is an East African corpus there is as yet no complete South African component, although Van Rooy (2005, 2006; Van Rooy and Schäfer 2003) has compiled the Tswana Learner English Corpus (TLEC) that forms part of the International Corpus of Learner English (ICLE). Currently, the following sub-corpora of ICE are available online:[10] Canada**,** Jamaica, Hong Kong, East Africa, India, Singapore, Philippines, Great Britain, New Zealand and Ireland.

In recent years, there has been a great deal of work using corpora in South Africa. Some of this has focused on issues of learner language, but the majority of the studies have been concerned with the nature and characteristics of black South African English (BSAE) as a variety of English, and the comparison of various South African varieties with each other and with varieties of English elsewhere in the world. Van Rooy (2013:10) observes that in South Africa, BSAE has received more attention than any other variety. His own Tswana Learner English Corpus (TLEC) project and De Klerk's (2002) corpus of spoken Xhosa-English are the best known corpora in South Africa to date and were both started in the early years of the millennium. Since then, other, smaller corpora have been collected, such as Pienaar's (Pienaar and De Klerk 2009) corpus of Indian South African English (ISAE) compiled for her master's project, and corpora compiled for studies such as the one discussed here.

Van Rooy's TLEC, which forms part of the ICLE and which has since been used in several studies (Henning 2006; Van Rooy 2005, 2006; Van Rooy and Schäfer 2003; Van Rooy and Terblanche 2006), was a collaborative project by Van Rooy and his colleagues at tertiary institutions in Mafikeng, Potchefstroom and

[10] http://ice-corpora.net/ice/avail.htm [accessed 18 February 2011]

Kimberley. It is made up of argumentative essays of between 400 and 1000 words written by students with a Tswana L1 background (Van Rooy 2006:46). Tswana is one of the nine official indigenous African languages in South Africa. This corpus is particularly relevant in the light of the present study in that it is made up of learner writing. In studies Van Rooy has conducted using this corpus (2005, 2006; Van Rooy and Schäfer 2003; Van Rooy and Terblanche 2006) he used this and other learner corpora from ICLE to investigate, for example, the extension of the progressive aspect in BSAE and the involved aspects of student writing.

The Xhosa-English Corpus (De Klerk 2002) is made up of spoken English transcribed from conversations between Xhosa-English speakers from the Eastern Cape. These conversations were completely unrehearsed, taking place face-to-face and recorded between 2000 and 2004 (De Klerk 2006b:128). The corpus contains about 540 000 running words of spoken English.

2.3.2 Corpora of learner language

Developments in corpus studies and a greater focus on learner language have led to the compilation of important corpora of learner language in the last few decades. One of the most influential today is the International Corpus of Learner English (ICLE) developed at the Université de Louvain in Belgium in 1990 under the leadership of Sylviane Granger. The ICLE comprises comparable corpora made up of writing in English by students from specific L1 backgrounds (McEnery and Hardie 2012:81–82). In order to be able to compare such writing to writing by native speakers of English, Granger and her team compiled the Louvain Corpus of Native English Essays (LOCNESS), made up of argumentative essays by British and American writers, and comprising about 324 000 words. McEnery and Hardie (2012:82) note that this work by Granger has had a dramatic impact on English corpus linguistics (ECL). A great deal of corpus research into learner language has come out of engagement with the ICLE and LOCNESS corpora, including some in South Africa, such as Henning (2006) and Van Rooy (2006) (mentioned in §2.3.1 above). These are discussed in §2.5.2 below.

Another important corpus of learner writing is the Hong Kong University of Science and Technology (HKUST) Learner Corpus, established by John Milton and made up of texts written in English by Chinese students and high school pupils with Cantonese as L1 (Pravec 2002). Flowerdew (2001) notes that various studies have been carried out using subsections of ICLE and HKUST, comparing non-native speaker or learner writing (or spoken English) with a parallel corpus of expert writing (the approach taken in this study) or with a larger reference corpus of expert writing, in order to determine for example which grammatical structures, lexis or discourse items are underused or overused by students in their academic writing (Flowerdew 2001:365). However, there are still not enough learner corpora available for research, with the result that, as in the case of the present study, many researchers are forced to compile their own. Most

corpora used in the development of English for Academic Purposes (EAP) materials, for instance, are made up of native speaker or expert writing or speech. Flowerdew suggests various reasons for the absence of learner corpora until fairly recently, but makes a strong argument for their usefulness in teaching:

> In spite of the difficulty of accessing learner corpora, I would still like to suggest that for materials to address students' needs and deficiencies fully, insights gleaned from learner corpora need to be employed to complement those from expert corpora for syllabus and materials design (Flowerdew 2001:364).

2.4 WHAT IS CORPUS LINGUISTICS?

Now that the concept of the corpus has been explained and some background to the development of corpora has been provided, this section returns to the central issue: what is corpus linguistics? McEnery and Hardie (2012:1) define corpus linguistics as 'an area which focuses upon a set of procedures, or methods, for studying language'. Tognini-Bonelli (2001:2) characterises corpus linguistics (CL) as 'an *empirical approach* to the description of *language use*; it operates within the framework of a *contextual* and *functional theory of meaning*; it makes use of *new technologies'* [author's emphasis]. Corpus linguistics has allowed researchers to explore theories of language that have emerged from findings based on actual language (McEnery and Hardie 2012:1). Thompson and Hunston (2006:8) believe that 'corpus linguistics is a methodology that can be aligned to any theoretical approach to language', and mention two main theories that have emerged from corpus linguistics: the theory that meaning is not located in single words but in what Sinclair (1991) calls 'units of meaning', and the theory that language is not simply a system into which lexical items are fitted, but rather a 'series of semi-fixed phrases' (Thompson and Hunston 2006:11–12).

Corpus linguistics is thus a method or an approach which is used in many branches of linguistics, such as lexicography, descriptive linguistics, applied linguistics, studies of language variation, dialect, register, style and more. The unifying factor in these diverse fields is the use of computerised corpora of texts or spontaneous speech (Léon 2007:1). There has been much argument – stretching from at least the time of a publication by Aarts and Meijs (1984 (Eds), cited in Taylor 2008:179), who first used the term 'corpus linguistics' in a book title, to the present day – on whether corpus linguistics is, for instance, a discipline, a methodology, a theory, a paradigm or all of these (Taylor 2008). Tognini-Bonelli (2001) and Mahlberg (2006) refer to corpus linguistics as a discipline. This has necessitated changes to the linguistic studies landscape: as Mahlberg (2006:370) observes, 'new descriptive tools are needed to account for the situation of real text, and ideas of theoretical frameworks to accommodate such tools have started to emerge'.

2.4.1 The starting point

The starting point of the discipline of corpus linguistics was 'observable data' (Tognini-Bonelli 2001:51), that is, the corpus. However, using a corpus (a collection, or literally a 'body') of language as the basis for the description of language is not a new phenomenon: historical linguistics has always been corpus-based and Tognini-Bonelli (2001:50) notes that as far back as a century ago, the study of language was equated with the observation of data. At the beginning of the twentieth century, however, the focus of modern linguistics moved from a data-based approach to one based on intuition and introspection (Biber and Finegan 1991). Although in the United States a decade or so later, Bloomfield (influential in American linguistics from the 1930s to the 1960s) rejected this and proposed a return to 'observation of normal speech' (Stubbs 1993:8), much of the data he and others following his lead studied were made up of invented sentences, and it was common in the studies of that time to discuss only very small numbers of such sentences.

With Chomsky in the late 1950s, however, came a return to the rejection of observable data as the basis for linguistic statements. He believed that a large body of language in use, a corpus, could not be regarded as relevant to linguistic enquiry mainly because it could not be relied upon as evidence of how the language faculty is organised (Chomsky 1962, cited in Léon 2007:4). Chomsky believed that inductive discovery procedures from a corpus of observed data could only observe surface phenomena: descriptions obtained by such methods were limited to the data which had been collected, and could not lead to any insights into the nature of language (Léon 2007:1). He separated his theory of language from usage by proposing the contrasting notions of competence (what one can do in theory) and performance (what one can do in practice) – similar to Saussure's *langue/parole* and syntagmatic/paradigmatic dualism (Stubbs 1993:2). Chomsky believed that corpora could not be used to study competence and neither observed data nor inductive procedures from observed data could provide reliable information on the linguistic intuition of native speakers. For Chomskyans, the source of insights into the structural nature of language is introspection (Chomsky 1956, cited in Léon 2007:2).

At much the same time, scholars such as Firth, Halliday and Sinclair became very influential in the development of the British school of corpus linguistics. In contrast to the Chomskyan perspective, they focused on the use of authentic language. These developments are discussed in more detail in the following section.

2.4.2 Modern corpus linguistics

As has been pointed out by Stubbs (1993), at around the same time as Chomsky was establishing transformational generative grammar in the late 1950s, linguistics across the Atlantic was continuing to develop rather differently under the influence of Firth, Halliday and Sinclair. This work gave rise to

principles which became central to neo-Firthian linguistics, one of the most important of which for corpus linguistics was that 'language should be studied in attested, authentic instances of use, not as intuitive, invented sentences; the language should be studied as whole texts, not as isolated sentences or text fragments; and […] texts must be studied comparatively across corpora' (Stubbs 1993:2).

These linguists agreed that linguistics should have as its 'essential subject' the study of meaning, and that form and meaning are inseparable; in other words, 'lexis and syntax are interdependent' (Stubbs 1993:2). Francis (1993:143) echoes this belief: 'syntactic structures and lexical items (or strings of lexical items) are co-selected, and […] it is impossible to look at one independently of the other'. Thus, their interdependence means that syntax and lexis are ultimately inseparable. The best way to study meaning, then, is through the use of a body of authentic, continuous texts, or a corpus.

Modern corpus linguistics has its roots in British corpus-based research. Léon (2007) argues that, in order to legitimise their claim to be an autonomous and unified linguistic field in spite of their obvious heterogeneity, and to ensure their theoretical position, corpus linguists developed arguments against Chomskyan generative grammar. Stubbs (1993) and Léon (2007) believe there are in fact two camps in the origins of British corpus linguistics. There is the Firth-Halliday-Sinclair line of development and the Quirk-Leech line. Stubbs observes that while the SEU in the 1960s collected large corpora of written and spoken data, which were used as evidence for Quirk et al.'s (1985) *A comprehensive grammar of the English language*, 'the precise relation between the corpus data and grammatical description is very vague' (Stubbs 1993:9). Sinclair's work in lexicography, on the other hand, was always based on 'attested data' (Léon 2007:12).

Rotimi (2006), in his discussion of the birth of Hallidayan linguistics, explains how this developed out of the work of Firth, who stressed that meaning is inseparable from context, both social and cultural, and that words are not independent. As Firth (1957, cited in Moon 2008:244) put it:

> The complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously.

Halliday, hugely influential in the late twentieth century, was a student of Firth and developed what he learnt from him into Systemic Functional Linguistics (SFL). Halliday and his followers, often also referred to as the neo-Firthian linguists,[11] studied language 'in relation to the social interactions which language encodes and the cultures within which these social actions are embedded' (Rotimi 2006:157). The focus of

[11] A group of scholars including John Sinclair, Michael Hoey, Susan Hunston, Michael Stubbs and Elena Tognini-Bonelli, many of whom were associated with the University of Birmingham, where Sinclair was Professor of Modern English from 1965 to 2000 (McEnery and Hardie 2012:122).

SFL is on how people interact and use language to make meaning. It emphasises the syntagmatic aspect of lexis, that is, the tendency of individual lexical items to co-occur habitually – in other words, collocation. Halliday 'made claims for lexis as a linguistic level of language analysis' (Rotimi 2006:159). Halliday challenged the distinction between grammaticalness and acceptability which Chomsky had proposed, and asked whether non-occurrence or low occurrence in a corpus was due to chance or was evidence of the ungrammaticability or rarity of a structure. He believed that, as far as corpora are concerned, the linguistic system is inherently probabilistic – 'frequency in text is the instantiation of probability in grammar' (Halliday 1991:30; Halliday 1992, cited in Léon 2007:12–13). He also believed that there is no fundamental difference between lexis and grammar.

Evidence from larger corpora in the 1980s and 1990s threw the descriptions of language based on pre-corpus studies into question. In the late 1970s, Sinclair began to collect the first computerised corpus of spoken British English at Edinburgh University (Sinclair, Jones and Daley 1970, cited in Tognini-Bonelli 2001:51) (see §2.2.2.2 above). The first study in lexicography using corpora was the COBUILD project (McEnery and Hardie 2012:123; Sinclair 1987, cited in Tognini-Bonelli 2001:51) and this reflected Sinclair's attitude to language theory and descriptive methodology. He continued in the belief of loose boundaries between lexis and grammar, that is, that collocation is a pattern of co-occurrence between two words which often occur in proximity to each other but not necessarily next to or in a fixed order (McEnery and Hardie 2012:123), which is the view of collocation taken in the present study. Sinclair assumed that, rather than using isolated words in rule-governed sequences, speakers tend to use ready-made linguistic forms, pre-packaged chunks – a view that is central to the present study (Léon 2007). This led to his 'open and closed choice principle', which will be discussed in greater detail in Chapter 4.

Sinclair's work showed that words are interconnected, that meaning is derived from context, and that collocation is key to understanding meaning (Moon 2008:243). Sinclair's approach to lexis was empirical and, like Firth, he was interested primarily in patterning and the meaning of text patterns (Moon 2008:244).

Sinclair 'prioritises a method, or group of methods, and a kind of data rather than a theory' (Hunston and Francis 2000:14). This approach to linguistics became known as corpus linguistics, the investigation of language through the observation of 'large amounts of naturally-occurring, electronically-stored discourse, using software which selects, sorts, matches, counts and calculates' (Hunston and Francis 2000:15). The data which are used in this approach differ from those used with other methods of linguistic investigation in that they are authentic and were not selected on linguistic grounds. In other words, they were not selected because they illustrate some specific aspect of language, because if this were the case, as Sinclair (1991:100) points out, it is 'likely to highlight the unusual in English and perhaps miss some of the regular,

humdrum patterns'. Hunston and Francis (2000:15) believe that these features of the data could be regarded as the principles underlying corpus linguistics.

Generally, the focus in language investigation now falls on performance (rather than on competence) and description, and on qualitative as well as quantitative analysis (Granger 1998a:3), although as will be shown below (§2.4.3), different approaches to dealing with corpora have emerged since they first appeared (cf. McEnery and Gabrielatos 2006). The trend has become one of investigating language in discourse, mostly authentic language, gathered both from native speakers' speaking and writing and from that of learners from various language backgrounds. This allows for a more holistic and realistic view of how learners acquire language structures and also of how they use this language in real-life situations. Now, more than ever, lexis is seen as the central principle of language and it is now almost universally accepted that much of language 'occurs in a sequence of morphemes that are more or less fixed in form' (Hunston and Francis 2000:7).

McEnery and Gabrielatos (2006:34) note that although it has been debated whether corpus linguistics should be regarded as a theory or as a methodology, they and many other researchers (cf. Tognini-Bonelli 2001) generally agree that corpus linguistics is more than simply a methodology. They quote Leech (1991), who called it a 'new research enterprise' and 'a new philosophical approach to the subject'. Halliday (1993, cited in Tognini-Bonelli 2001:1) claims that corpus linguistics 'reunites the activities of data gathering and theorising, which has led to a qualitative change in our understanding of language'.

What is generally agreed is that corpus linguistics is empirical, as it examines authentic language use and draws conclusions from it rather than from intuitions. Tognini-Bonelli (2001:1) argues, too, that it has theoretical status because researchers observe language facts which lead to the formulation of hypotheses and generalisations, allowing linguists to use new parameters to explain data. These hypotheses and generalisations are then brought together in a theoretical statement. This theoretical status means that corpus linguistics can contribute to applications such as lexicography, language teaching, translation, stylistics – the list is a long one – but the present study confines itself to the use of corpus linguistics and corpora in the context of second language learning, and second language writing in particular.

In summarising this section, then, it can be said that the differences between Chomskyan assumptions and British neo-Firthian principles, the foundation of corpus linguistics as we know it today, are fundamental. Stated very briefly, these differences are, firstly, that Chomskyan linguists believe that linguistic study could be based on intuitive data and isolated sentences, while neo-Firthian linguists believe that language should be studied in actual attested and authentic instances of use; and secondly, that Chomskyans believe that grammar is autonomous and independent of meaning (Chomsky 1957, cited in Stubbs 1993:13), while neo-

Firthians believe that linguistics is concerned with the study of meaning; hence, form and meaning are inseparable.

One of the most crucial principles on which neo-Firthian linguistics was premised and one that is central to the present study is that linguistic behaviour demonstrates that 'language in use is a balance between "routine and creation", and that much of our language use is routine' (Stubbs 1993:2).

2.4.3 Debates in corpus linguistics

So far, this chapter has focused mainly on the neo-Firthian approach to corpus linguistics but, as McEnery and Hardie (2012:147–149) point out (§2.4), this is not the approach followed by all corpus linguists today. Tognini-Bonelli (2001:177), taking her cue from Granger's (1998a) notion of a continuum of corpus approaches, argued for the establishment of a 'new discipline' within linguistics, and within corpus linguistics itself. She suggested the name 'corpus-driven linguistics' (CDL), in contrast with 'corpus-based linguistics' (CBL). These approaches are termed, perhaps more usefully, by McEnery and Hardie (2012:147) as 'corpus-as-theory' and 'corpus-as-method' respectively.

According to the neo-Firthians, in corpus-based research, or corpus-as-method, data (that is, the corpus) are used to test or provide examples for the theories which were formulated before large corpora became available to inform the study of language (Tognini-Bonelli 2001). Römer (2005) believes that corpus-based linguists, or those who follow the approach of corpus-as-method, 'do not put the corpus at the centre of their research but see it as a welcome tool which provides them with frequency data, attested illustrative examples, or with answers to questions of grammaticality or acceptability' (2005:9). In other words, in corpus-based linguistics (CBL), corpora are used as instruments together with other strategies and other data. On the other hand, in corpus-driven linguistics (CDL), or a corpus-as-theory approach, the linguist attempts to explore the language, making no distinction between lexis and grammar, 'free from the influence of existing theoretical frameworks, which are considered to be based on intuitions' (McEnery and Gabrielatos 2006:36), and for this reason, unreliable. It is for these reasons that corpus-based linguists tend to use annotated corpora while corpus-driven linguists do not.

Annotation allows the researcher to easily quantify categories, such as parts of speech or semantic roles. McEnery and Wilson (1996:32) believe that the annotation of a corpus increases its 'utility'. However, categorisation of the data into word classes sometimes results in the meaning of certain lexical items being obscured (Römer 2005:9). Annotation is conceived by corpus-driven linguists as being influenced both by theory and by the intuitions of the individual annotating a corpus. They believe, as Römer (2005:10) expresses it, that annotation does not necessarily make the data any richer. Corpus-driven linguists allow the data to do the talking, as it were (Granger 1998a), and search by word forms rather than by annotated

categories. 'Findings are directly derived from the data; no filtering through existing concepts is supposed to take place' (Römer 2005:10).

CDL tends to avoid, though not entirely ignore, intuition; the evaluation of the data does demand subjective judgements from the researcher which are based on experience as a language user and learner (Römer 2005:10), but the evidence from the data should always come before intuitions. Römer (2005:10–11) believes that taking a corpus-driven approach implies that the evidence cannot ever be ignored; instead, it must be 'accepted and reflected in new theories. New theoretical statements are derived from the data and have to be fully consistent with the corpus evidence'. McEnery and Hardie (2012:148–149) sum up the differences in these approaches as follows:

> While corpus-as-theory rejects any explanation of language patterns that does not derive from the analyst's interaction with the data, corpus-as-method considers corpora and corpus techniques to be sources of empirical data that may be deployed in support or refutation of any explanatory theory about language – even a theory devised in whole or in part without reference to corpus data.

What is clear, though, is that these dichotomies are not clear-cut. McEnery and Hardie (2012:150) point out that the term 'corpus-driven' has been used to apply to any 'inductive, bottom-up research using raw corpus data', regardless of whether the researcher subscribes to the neo-Firthian theoretical stance. They mention that Biber (2009) uses this term for his approach although it is obvious that he does not follow neo-Firthian theories. Other sources such as Gries (2010) and Gilquin and Gries (2009) also use this term in a somewhat different way from Tognini-Bonelli (2001). Gilquin and Gries (2009:10) describe the corpus-driven approach as typical of linguists who 'approach corpus data in an exploratory fashion, i.e. without rigorously formulated hypotheses'.

In the final analysis, McEnery and Hardie (2012) believe that the terms CBL and CDL are really misleading in that they imply that the main difference between the two approaches is the extent to which researchers rely on the empirical evidence from the corpus, when in fact 'respect for the empirical evidence of the corpus is probably one of the closest points of agreement between the two traditions of corpus linguistics' (McEnery and Hardie 2012:151). In fact, they believe that the distinction between the two approaches implies a 'sliding scale' (McEnery and Hardie 2012:151) or as Granger (1998a) expresses it, that these approaches should not be seen as polar opposites, but rather as the two end points on a continuum. At the corpus-based end, the analysis starts from a hypothesis based on the research literature on, for example, second language acquisition (SLA) research and uses a corpus to test this hypothesis (Granger 1998a:15). At the other end of the continuum is the corpus-driven approach, sometimes referred to as a 'hypothesis finding' approach (Scholfield 1995, cited in Granger 1998a:15): recurrent patterns and frequency distributions observed in the corpus data are used to form insights about language, without the influence

of pre-existing theories and frameworks and with the purpose of developing a purely empirical theory (McEnery and Gabrielatos 2006:36). In the end, perhaps what best separates these two schools of thought is their stance on the conceptual status of the corpus and of corpus linguistics – as having theoretical status versus being regarded as a linguistic methodology (McEnery and Hardie 2012:151).

According to the definitions provided above, the present study can best be defined as a hybrid (Biber 2009:281), something which Biber believes is quite acceptable. Tognini-Bonelli (2001:87) explains that the corpus-driven approach 'aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context'. Römer (2005:7) puts it a little more simply, perhaps: a study which takes the corpus-driven approach 'is highly committed to the data it starts from and [...] tries to derive observational and theoretical findings from there, always trying not to lose contact with the corpora'. However, as this study starts with a particular lexico-grammatical structure in mind, and selects three verbs with which to investigate the structure, it can be said to take a corpus-based approach initially, or a corpus-as-method approach, in that I have used the data to provide me with examples of these verbs as they occur in MWUs.

2.5 SOME CORPUS RESEARCH RELEVANT TO STUDENT WRITING

Phase 2 of this study is concerned with the use of computerised language corpora (CLCs) to investigate an aspect of the depth of students' vocabulary knowledge, their use of MWUs containing the delexical verbs *have*, *make* and *take* in their writing, and to compare students' use of these combinations to that of expert writers. Various studies have been conducted specifically in the area of learner[12] writing, both abroad and in South Africa. Some of these are mentioned in the following sections, while Chapter 4 focuses on those studies which relate specifically to learners' use of MWUs and delexical verbs.

Computerised language corpora (CLCs) allow access to learners' interlanguage as a whole, not just to their errors (Granger 1998a:6). Corpora have other advantages too; for instance, they allow materials designers to prioritise the grammar and vocabulary that is most frequently used (Hunston 2002) and corpora provide researchers, test designers and teachers with examples of authentic language (Jaén 2007). In this way, too, learner corpora offer potential insights for both syllabus design and methodology (Liu and Shaw 2001). Biber and Conrad (2001:332) make two generalisations from their findings from corpora which they believe are vital to ESL and EFL teaching. Firstly, quantitative corpus analyses have revealed that register is of paramount importance – linguists must consider patterns of use across registers if they are to provide a complete analysis of grammatical patterns. Secondly, they confirm the generally held view that intuition is

[12] The students in this study are all 'learners' in the sense that they can be regarded as 'novice' writers in the context of university student academic writing, but the group includes speakers of English as a first language, as well as those for whom it is not. Thus they are not all 'learners' of the language in the sense that Granger or others apply the term (i.e. to second or foreign learners).

unreliable (cf. McEnery and Gabrielatos 2006), noting that corpus studies show that intuitions about use are often incorrect.

2.5.1 International corpus studies

Flowerdew (2001) notes that various studies have been carried out using subsections of the ICLE and HKUST Learner Corpus (see §2.3.2), comparing non-native speaker or learner writing (or spoken English) with a parallel corpus of expert writing (as is the case in this study) or with a larger reference corpus of expert writing, in order to determine aspects such as which grammatical structures, lexis or discourse items are underused or overused by students in their academic writing (Flowerdew 2001:365).

Flowerdew (2001) describes a study she did using the HKUST Learner Corpus. She based it on an earlier study by Pickard (1994, cited in Flowerdew 2001), who examined the vocabulary used in reports by second and third-year undergraduate students in a Technical Communications Skills course. Flowerdew focused on key items such as *findings*, *research*, *survey*, *questionnaire*, *respondents* and so forth. Using the KWIC (key-word-in-context) tool, her findings were that although students knew and used the key vocabulary, they were not aware of the lexico-grammatical environment in which these words usually occur. This echoes research by Howarth (1998a, b), which is described in greater detail in the following chapters. He drew up a collocational framework and found that the restricted collocation category, including such combinations as *make a claim* or *take a decision*, was the one which caused learners the most difficulties. This has reference to the present study, where the MWUs in question can be regarded as restricted collocations as described by Howarth (1998a) (see Chapter 4).

Contrastive studies have also been done by Granger (1998b, discussed in Chapter 4), the founder of the ICLE project, using the French component of the ICLE corpus and the Louvain Corpus of Native English Essays (LOCNESS) (see §2.3.2) as the control corpus. Granger and Rayson (1998), for instance, examined patterns of over- and underuse of various word categories, such as articles, determiners, pronouns and prepositions in the learner corpus, and found that there was a 'significant overuse of first and second personal pronouns, but an underuse of prepositions, a category often associated with nominalisations in informative academic writing' (Flowerdew 2001:365). These and other findings led these researchers to conclude that the learner data in their study had many of the stylistic features of spoken English, but 'practically none of the features typical of academic writing' (Granger and Rayson 1998:129). This is an aspect that is discussed in greater detail with reference to the present study in Chapter 4.

Similar studies making use of learner corpora include those by Granger and Tyson (1996) on connector usage, by Granger (1998b) on formulaic sentence builders, by Altenberg and Granger (2001) on the lexical and grammatical patterning of the verb *make*, by Nesselhauf (2005) on collocations in her corpus of texts

by German advanced learners of English and by Kaszubski (2000) on aspects of phraseology of high-frequency verbs among Polish learners of English. These and other studies are discussed in detail in Chapter 4.

2.5.2 Corpus studies in South Africa

As mentioned above (§2.3.1), in recent years a great deal of work has been done in South Africa using corpora. Most of these studies have investigated characteristics of BSAE as a variety of English, or have compared various South African varieties with each other and with varieties of English elsewhere in the world.

In South Africa, a number of studies have been conducted using the ICLE with the LOCNESS as the control corpus. Van Rooy has conducted several studies (Van Rooy 2005, 2006; Van Rooy and Terblanche 2006) using the Tswana Learner English Corpus (TLEC) which forms part of the ICLE (see §2.3.1). In a study with Terblanche (Van Rooy and Terblanche 2006), the authors used the TLEC and LOCNESS corpora to 'determine and explain a coherent set of characteristic features of student writing produced by second language users of English in South Africa who had their secondary education in township schools' (2006:162). They began the study with the perception that students' writing has several conversational features which are out of place in academic writing. As in the present study, they used the software WordSmith Tools in their analysis.

Their findings revealed that the student writing in the TLEC was indeed more informal and colloquial than that in the LOCNESS and had more conversational features such as reduction phenomena. Information tended to be more fragmented and writers were more cautious in the claims they made. They also found that cohesive devices were used in a more general way by these students and were more likely to be ambiguous than in the writing in the control corpus (Van Rooy and Terblanche 2006:175). But, contrary to generally held views, the two corpora were more alike than they had expected and the differences between them were not as pronounced as the differences between student writing, regardless of first language, and other registers such as academic writing, personal letters, face-to-face conversation and general fiction (Van Rooy and Terblanche 2006:175–176). This last point is particularly relevant to the present study, which investigates writing by students from several language backgrounds in South Africa, but which does not focus specifically on BSAE, as many South African studies have tended to do.

In another South African study, Henning (2006) investigated linking adverbials and their influence on cohesion and coherence in the academic writing of English native speakers (what she calls ENL), ESL and EFL speakers. She used student texts from the LOCNESS corpus (her ENL sample), texts from the Tswana Learner English corpus (TLEC) as the ESL sample, and the Dutch English component of the ICLE as the EFL

component. She adopted a corpus-based approach, using a tagger developed by Henning and Badenhorst (2004, cited in Henning 2006:75). She found that the writing by the South African students (TLEC) revealed fewer linking adverbials than the writers in the LOCNESS corpus, which she interpreted as a gap in the language proficiency of the BSAE speakers. In her study, the ESL data differed from both the ENL and the EFL data, leading her to claim that BSAE differs 'markedly' from Standard English (Henning 2006:127).

As noted above (§2.3.1), De Klerk, then at Rhodes University in Grahamstown, did a great deal of corpus work among Xhosa-English speakers in the Eastern Cape. She subsequently compiled a corpus of spontaneous English spoken by Xhosa L1 speakers (De Klerk 2002, 2006a, b) and used this in several studies. De Klerk (2006b) used this corpus of 'spontaneous' English and a smaller corpus of teacher-talk in English by mother-tongue Xhosa teachers in classrooms in Grahamstown to compare language use in the two cases. She was particularly interested in how the discourse patterns of these teachers reflected those used by L1 teachers. She found that there were differences between the 'patterns of usage' in the Xhosa-English Corpus and those used by the teachers in the classroom, what she refers to as the 'gatekeepers' (2006b:133). She found that the teachers mostly used Standard English with very few of the distinctive features of the Xhosa-English conversations: the features which were shared by the two corpora included the high usage of progressives and of *can be able*. She ascribed these findings to the fact that Grahamstown is a small university city and an educational centre; teachers were more likely than their rural counterparts to come into regular contact with L1 Standard English speakers and could thus be expected to use a variety of English that was closer to the standard.

In a study of a sub-variety of South African English (SAE), Pienaar and De Klerk (2009) describe the compilation of a corpus of Indian South African English (ISAE). This is a 60 000-word corpus of 30 speech samples, each made up of about 2000 running words. They decided to use speech and not written texts because ISAE is chiefly an oral dialect (Pienaar and De Klerk 2009:359). The data were gathered from 49 South African-born Indians. In preliminary investigations using WordSmith Tools they were able to confirm that features which had been identified by other researchers in this sub-variety of SAE existed in their own corpus (2009:366).

BSAE has generated interest beyond our borders, too. Minow (2010), a German scholar, collected a corpus of speech from 45 informants, mostly L1 Xhosa speakers and some L1 Tswana and Southern Sotho speakers. The aim of her study was to identify which features of BSAE, identified by researchers but not always quantified, had become stable features of the variety. She focused on four features of BSAE (past tense marking, progressive aspect usage, article omission and left dislocation), analysing her data manually. Her findings led her to believe that, despite the small number of speakers in her sample, there was evidence for several BSA Englishes, not just one, a claim which De Klerk (2002:25) supports.

It would seem that there is scope for a great deal of corpus research in the South African context, and it is hoped that the present study will add to this body of knowledge.

2.6 CONCLUSION

This chapter set out to review what the literature has revealed about the advent and development of corpus linguistics and its significance in the field of linguistic, and particularly vocabulary, studies. As Pienaar and De Klerk (2009:354) observe, 'the use of a corpus has become the sine qua non in many areas of linguistic enquiry', a sentiment echoed by the scholars mentioned in this chapter. In a sense, corpus linguistics has revolutionised vocabulary studies, and in particular the study of words and their relationship with each other in context. Corpora today are computerised and huge amounts of data can easily be submitted to a range of software tools, providing a quantitative approach to learner language which was far more difficult to achieve before the advent of CLCs and the tools that have been developed to analyse them. Studies are now easier to replicate and more likely to be statistically reliable (Granger 1998a:3; McEnery and Gabrielatos 2006:34). This type of research includes quantitative information – measures such as frequency counts and statistical measures of strength of lexical co-occurrences – as well as the qualitative analysis of word patterns in different sentence contexts. Qualitative analysis of this nature includes the manual investigation of concordance lines, the identification, extraction and exploration of patterns and the classification of particular words or groups of words in the corpus, among others.

This chapter has dealt with the development of the discipline of corpus linguistics and has focused particularly on work done with learner corpora. In the following chapter, the focus is on vocabulary – research in this field of linguistic studies and its importance to academic proficiency.

Chapter 3

Vocabulary and vocabulary studies

> Vocabulary is an essential building block of language and, as such, it makes sense to be able to measure learners' knowledge of it. (Schmitt, Schmitt and Clapham 2001:55)

3.1 INTRODUCTION

The number of words that scholars believe a learner needs to be successful in academic study has varied widely and consensus has yet to be reached, partly because of the fluid nature of researchers' definitions of a word. As Vermeer observes, '[t]he figures reported in the literature are largely determined by what is meant by a word' (Vermeer 2001:220) and Milton and Treffers-Daller (2013:154) caution that 'estimates of vocabulary size need to be treated with great caution because of the methodologies involved'. Vermeer also notes that there are several degrees to 'knowing' a word: 'Knowing or not knowing words may seem like a dichotomous distinction, but there is, in fact, a continuum ranging from not knowing, to recognizing, to knowing roughly, to describing very accurately' (2001:221). Since the 1980s, many scholars have tried to quantify the vocabulary needed to read competently and with ease: researchers such as Hazenberg and Hulstijn (1996), Schmitt et al. (2001), Schmitt and Zimmerman (2002), Morris and Cobb (2004) and Milton and Treffers-Daller (2011, 2013) have all made suggestions about the size of vocabulary required for proficient reading, and by extension then, success in academic studies.

In this study, the investigation of students' vocabulary is approached from the perspective of both discrete item knowledge (breadth or size of vocabulary) and depth of word knowledge, that is, the degree of knowledge of particular words, specifically, how certain high-frequency words can be used in certain types of multiword units (MWUs). In Phase 1, the VLT (Laufer and Nation 1995) is used to measure the number of words known productively; in Phase 2 an aspect of the depth of knowledge of a particular type of word is assessed with the help of corpus analysis. Phase 3 brings these two aspects of word knowledge together in an investigation of the links between word knowledge and the use of MWUs, and academic performance. In this phase, the study examines the relationships between various aspects of discrete and MWU vocabulary knowledge on the one hand and academic performance on the other. The focus of the study moves thus from individual word knowledge to knowledge of MWUs, bringing the two together in the third phase.

In this chapter, a brief overview of research which has been conducted in the area of vocabulary studies is provided, particularly studies on the vocabulary students need to succeed in academic study, that is, the number of words they need to know, the type of words and the type of knowledge of these words. This necessarily draws on research in the fields of reading as well as writing, given that studies have long shown that vocabulary knowledge is vital to proficiency in both. This chapter focuses thus on what research has revealed about foundational knowledge of high-frequency words on the one hand, and knowledge of lower frequency academic vocabulary on the other.

The following section presents a discussion of what research has found about the importance of size or breadth of vocabulary. The chapter moves on to a discussion of the concept of depth of vocabulary knowledge, and this is followed by a section which deals with word lists and academic vocabulary. The final section of this chapter discusses the relationship between vocabulary and academic performance, both at school and at university level.

3.2 ESTABLISHING THE SIZE OF STUDENTS' VOCABULARY

Research has long indicated a relationship between vocabulary size and the ability to use language in various ways. Size implies measurement, and a common measure in vocabulary studies is the word, or more specifically, the word family. In this study, 'word' (unless otherwise specified) refers to a word family or lemma, comprising a base word and both its inflected and derived forms (Qian 2002; Read 2004, 2007; Schmitt and Zimmerman 2002). For example, the word family or lemma *HAVE* includes all the derived forms of this word such as *have*, *has* and *had*. Knowledge of the most frequent 2000 words in English provides the bulk of the lexical resources required for basic everyday oral communication (Schonell et al. 1956, cited in Schmitt et al. 2001:55). The next 1000 words (3000-word level) are useful for spoken discourse but, more importantly perhaps, this level of knowledge is the threshold at which beginning readers can start to learn from context (Nation and Waring 1997:11) and at which second language learners are able to transfer language skills learnt in their L1 to their L2 reading and to begin to read authentic texts (Cooper 2000; Schmitt et al. 2001:56). Laufer (1997) confirms this in her findings that a vocabulary of 3000 word families, or about 5000 words, is necessary for general reading comprehension and should allow readers coverage of 90 to 95% of the running words, or tokens, in a text. This is supported by Schmitt et al. (2001:56), who note that most research has suggested that knowledge of the most frequent 5000 words should provide learners with enough vocabulary to read authentic texts. However, and more importantly for this study, it is commonly believed that for learners to be able to achieve a 95% coverage of academic texts they need to have mastered far more than the 2000 high-frequency words of a language (core vocabulary): they need also to have knowledge of some academic words and technical terms (Paquot 2010:9–10).

3.2.1 Measuring breadth: the Vocabulary Levels Test

The VLT (see Appendix A, where the full test is provided) was designed by Nation (1983) 'to provide an estimate of vocabulary size for second language (L2) learners of general or academic English' (Schmitt et al. 2001:56). Most instruments designed to test vocabulary make a distinction between receptive (passive) and productive (active) vocabulary. Laufer (1994) defines productive knowledge as what a learner needs to know about a word to use it in speaking and writing. Although reading is not a passive activity, and readers participate actively in the reading process, it is generally accepted that more knowledge is required for productive language performance (Nation 1990:31), because productive usage means that learners must understand the connotations as well as the denotations of the words they use, whereas reading and listening do not always demand such detailed or specific knowledge. Laufer and Nation (1995) identify two types of productive vocabulary. They use the term 'controlled productive ability' to refer to the ability to use a word when forced to do so by a teacher or researcher, for instance in a restricted context such as a fill-in task where a sentence context is provided and the missing target word has to be supplied. The second type they term 'free productive ability', that is, the use of a word in an unrestricted context such as a sentence-writing task.

The VLT (Nation 1983, 1990) assumes that the vocabulary of English or any language consists of 'a series of levels based on frequency of occurrence' (Laufer and Nation 1999:35). This test provides a useful way of estimating a learner's vocabulary size at each of the four frequency levels (2000-, 3000-, 5000- and 10 000-word levels) as well as academic vocabulary, based on the University Word List[13] (Xue and Nation 1984, discussed in more detail in §3.4 below). Because of this variation in frequencies and therefore usage, a distinction is made in most vocabulary studies between the high-frequency words represented by the most frequent 2000 words (Nation and Hwang 1995, in Laufer and Nation 1999:35; Paquot 2010; Read 2004) and the large number of low-frequency words. Paquot (2010:10) defines high-frequency words, what she refers to as 'core vocabulary', as 'high-frequency in most languages and including function and content words'. Stubbs (1986:104) calls such core words 'pragmatically neutral' or 'nuclear vocabulary', because they have no particular connotations and no associations with specific contexts, nor are they found only in specific types of discourse or exclusively in formal language or informal language. They occur in both spoken and written language and have 'no necessary attitudinal, emotional or evaluative connotations' (Stubbs 1986:104).

The distinction between high- and low-frequency words was the motivation behind the construction of the original VLT and its productive version: to allow teachers to determine the stage where their learners were in their vocabulary development, and to help them decide what they needed to teach (Laufer and Nation

[13] This list provides coverage of 87% of words in the university level academic texts that these authors included in their study. It is made up of 808 words grouped into 11 frequency levels.

1999:36). The most frequent 10 words account for about 25% of the tokens in spoken and written language. Altogether, the 1000 most frequent words account for about 75% of the running words in formal written texts and about 84% of informal spoken use. On the other hand, the least frequent words (the 10 000-word level, that is the 5000 words between 5001 and 10 000) make up less than 1% of the running words in a text (Laufer and Nation 1999:35). It should be noted here that these estimates vary from researcher to researcher – as noted below, Hirsch and Nation (1992) arrived at slightly different figures, as did Read (2004) – but these figures are all fairly similar.

The list of high-frequency, core words that is still most commonly used today is the General Service List (GSL) compiled by West (1953). This contains about 2000 word families and was compiled from a five million-word corpus of written English. The GSL has been tested in several studies and has been found to provide coverage of up to 92% of fiction texts (Hirsch and Nation 1992:691) and up to 68% of academic texts (Coxhead 2000). Although the list is criticised by some for being old-fashioned, and containing words that today have only a limited use (Paquot (2010:11) mentions words such as *crown*, *coal*, *ornament* and *vessel* in this regard), and Engels (1968, cited in Paquot 2010:11) found that the second 1000 word families in the GSL provided under 10% coverage of the 10 texts of 1000 tokens that he analysed, many researchers still use it and believe that it provides useful information (Nation and Waring 1997:13) and it is the basis of the first level of the VLT.

As Read (2004:148) observes, only a small proportion of the words in the English language make up a relatively huge proportion of the running words in any text. It therefore makes sense that learners should concentrate at first on learning the 2000 most frequent words which account for at least 80% of running words in a text (Read 2004:148). Thus, a distinction in teaching is usually made on the basis of cost and benefit – the cost of the time and effort spent on teaching these words, and the benefit of the number of opportunities the learner is likely to have to use them (Laufer and Nation 1999:36). Laufer and Nation (1999) cite the VLT as useful to teachers when deciding what aspects of vocabulary to focus on with particular groups, particularly because teaching low-frequency vocabulary requires different approaches to those used to teach high-frequency vocabulary.

Phase 1 of this study starts with a measurement of the size of students' productive vocabulary, part of establishing novice writers' proficiency in written English, by assessing at the outset how many words they know (can recognise and use productively, in this case in a controlled productive test). To this end, a version of the VLT, the controlled-production vocabulary-levels test, that is, the *Active Version of the Vocabulary Levels Test* (Laufer and Nation 1995), which is freely available to researchers on the internet, is used. Laufer and Nation (1999:33–34) found this test to be 'reliable, valid (in that all levels distinguished

between proficiency groups) and practical' and agree that information about specific aspects of the proficiency of language learners can be useful for diagnostic purposes and in curriculum design.

To sum up, and notwithstanding the various views on the size of vocabulary needed to speak fluently, listen well, read generally, and read academic texts with understanding (discussed in more detail below in §3.5.1 and §3.5.2), the VLT is a test which many researchers have found useful, valid and reliable to use, and above all, practical. It has been a most influential instrument and has been used regularly from its emergence in the early 1980s to the present day, with several new versions being compiled by later researchers such as Schmitt et al. (2001). For these reasons it was used in the first phase of this study.

The following section discusses the concept of depth of vocabulary knowledge, that is, the kind of knowledge of a word which learners need to have.

3.3 DEPTH OF VOCABULARY KNOWLEDGE

Scholars agree that there are several aspects to the notion of 'knowing a word' and different degrees of this knowledge. Testing the breadth of a learner's vocabulary tests the learner's ability to identify the written form of a word with a simple statement of its associated meaning; depth involves knowing much more about a word if it is to become part of a learner's functional lexicon (Read 2007:113). Such knowledge includes knowledge of collocations, word associations, how to use words in context, and their related meanings (Laufer and Nation 1999:34), such as derivations and inflected forms, and of course today the dimension which Hancioğlu et al. (2008:460) call lexico-knowledge; the complex 'interlocking of systems and levels' because 'really "knowing" a word means knowing a lot of other words' (Hancioğlu et al. 2008:474). Read (2004:155) observes that

> depth of knowledge focuses on the idea that for useful higher-frequency words learners need to have more than just a superficial understanding of the meaning; they should develop a rich and specific meaning representation as well as knowledge of the word's formal features, syntactic functioning, collocational possibilities, register characteristics, and so on.

In addition to knowledge of the most frequent words, however, learners need to be able to use lower frequency words appropriately in academic productive tasks (Hancioğlu et al. 2008:461) and this means having some depth of knowledge about these words. Knowledge of connotations allows advanced readers to make judgements about diction and register, vital aspects of fully understanding an academic text. Hancioğlu et al. (2008:461) believe that for learners of a language, 'issues of breadth and depth become even more critical when we consider productive skills' (2008:461). This point is particularly pertinent to the present study, given its focus in Phases 2 and 3 on vocabulary use in academic writing.

Thus, as Laufer and Paribakht (1998:367) express it, lexical knowledge can be seen as 'a continuum consisting of several levels and dimensions of knowledge', moving from a 'vague familiarity' with a word when one meets it, to recognising it when it is seen or heard, to being able 'to use the word correctly in free production'.

3.3.1 Measuring depth of word knowledge

The trend in research on depth of vocabulary knowledge has been for researchers to focus on specific aspects of word knowledge, possibly because it is very difficult to measure all aspects of what it is to 'know' a word. Studies have dealt with aspects such as collocational knowledge (Bahns and Eldaw 1993; Nesselhauf 2003, 2005; Qian 2002), derivational knowledge (Schmitt and Zimmerman 2002) and knowledge of polysemy and synonymy (Qian 2002). Shin and Nation (2008) (see §3.4) have substantiated the importance of collocational knowledge, for instance, through their study of collocations of the 1000 most frequent spoken word types in the BNC in order to establish those collocations which would be most useful for learners in an elementary conversational English course (2008:339).

Laufer (1998:257) suggests a 'multiple test approach' to testing vocabulary, as there is no single test to measure both vocabulary size and depth: different instruments are required to test different aspects of word knowledge (Laufer and Nation 1999; Qian 2002; Qing 2009; Read 2004). For instance, the VLT treats vocabulary as a trait, a mental characteristic of a learner and one which can be described and measured without the words being used in any particular context. Corpus investigations, on the other hand, are founded on an interactionist view of vocabulary, one which assesses learners' ability to use the words they know appropriately in particular contexts (Read 2007:115). As noted in the section above, it is often difficult to entirely separate breadth and depth, and most researchers see the value of testing both dimensions of word knowledge. Vermeer (2001) asserts that breadth and depth are closely related and that the greater one's vocabulary, the deeper one's vocabulary knowledge, and vice versa. She observes crucially that 'knowledge of words is now considered the most important factor in language proficiency and school success – in part due to its close ties with text comprehension' (Vermeer 2001:217). Qian (2002) believes that it is important that both aspects, breadth and depth, are kept in mind as both are important for reading comprehension. He is referring to university students, but Vermeer's study, conducted as it was with Dutch L1 and L2 children between the ages of four and seven, reminds us that the foundations of vocabulary knowledge must be laid in pre-school and primary education.

In his study, Qian (2002:520) set out to establish the contribution of vocabulary size and three aspects of vocabulary depth (synonymy, polysemy, and collocation) to basic reading comprehension (2002:520), using a sample of ESL students from 19 language backgrounds at the intermediate level at a Canadian university. Qian used four measures for this purpose: the *Reading for Basic Comprehension Measure,* a version of the

TOEFL reading comprehension subtest, as the criterion measure of reading for basic comprehension (TOEFL-RBC) (Qian 2002:523), and three vocabulary measures as the independent predictor variables: the *Depth-of-Vocabulary-Knowledge Measure* (DVK), which tests depth of receptive English vocabulary knowledge by measuring three elements of vocabulary: synonymy, polysemy, and collocation (Qian 2002:524); the *Vocabulary Levels Test* (labelled VS) (Nation 1983), which measures vocabulary size by testing single meanings of content words; and the *TOEFL Vocabulary Item Measure* (TOEFL-VIM), which measures knowledge of English synonyms in a limited context, and contains 30 discrete vocabulary items developed for TOEFL administrations before July 1995.

Qian (2002:531) found that scores on all three predictors, that is, DVK, VS (the VLT) and TOEFL-VIM, were 'highly intercorrelated' and although he found all three vocabulary measures to be equally useful in predicting reading performance, the DVK was superior in that it provided a greater 'positive washback' (Qian 2002:532) effect on teaching and learning of new vocabulary because it assessed knowledge of polysemy and collocation rather than of single meanings of target words. He also concluded that using measures of depth and size of vocabulary together, rather than one or the other on their own, would increase the ability to predict reading performance (Qian 2002:532). Qian thus concluded that vocabulary assessment could be enhanced by using questions based on both vocabulary size and vocabulary depth.

In their study, Schmitt and Zimmerman (2002:146) expand on the notion of vocabulary knowledge as a continuum (Laufer and Paribakht 1998; Vermeer 2001). Vocabulary knowledge is incremental because there are so many components to be mastered: 'it would be impossible to learn all of these components fully from only one exposure to a word' (Schmitt and Zimmerman 2002:146). For instance, Schmitt and Zimmerman (2002) found that although ESL university students had a partial knowledge of the derived forms of words, this knowledge could not be taken for granted, suggesting that scholars may actually underestimate the number of words that must be learnt for successful reading of academic texts. Using a version of the Test of Academic Lexicon (TAL), Schmitt and Zimmerman (2002) investigated students' productive derivational knowledge of members of word families in relation to more global knowledge of target words. Their subjects were English NS and ESL students from universities in the US and the UK. Selecting target words and word families from Coxhead's (2000) Academic Word List of about 850 academic words, with which they expected their students to be familiar, Schmitt and Zimmerman (2002) measured their participants' depth of lexical knowledge by eliciting students' ratings of their own word knowledge. They then measured their students' productive knowledge of derivative forms by asking them to provide the appropriate derivative form of the target word, or to indicate that no derivative existed (Schmitt and Zimmerman 2002:153–154). Their findings suggested that global mastery of derivative forms may increase with general proficiency, although even very advanced users of English were unlikely to know all derivations of a target word (Schmitt and Zimmerman 2002:162). They found that, like the knowledge of vocabulary in general, derivational knowledge appears to be incremental, and learners typically know some

but not all the derivative forms making up a word family. As was perhaps to be expected, the ESL undergraduates in their study knew the lowest number of derivative forms, the proficient ESL graduate students knew more than this group and the NS students knew the highest number of all.

The reliability of Schmitt and Zimmerman's (2002) findings, based as they were on students' self-ratings, could of course be questioned; self-rating exercises and questionnaires are notoriously flawed as far as the reliability of their results is concerned. In a South African study, Coetzee-Van Rooy (2011) (see §1.2) found that students tended to have an inflated view of their own proficiency in English and did not believe that they had any problems in this regard, rating their proficiency as good, when their actual performance in the language belied this. Stephan et al. (2004:42) also observe that many South African university students believe that they have few difficulties with English even though their lecturers see English as the main cause of their academic difficulties. Studies by Oyetunji (2011) and Lukhele (2010) also found that students' self-assessments were inaccurate. In Oyetunji's study, while respondents claimed that they made use of reading strategies, their performance did not bear this out (2011:155). Lukhele (2010) found that although the questionnaire scores indicated that her subjects appeared to have positive attitudes to reading, their performance in the reading and vocabulary tests 'fell short of expectations' (2011:152–3).

Nonetheless, Schmitt and Zimmerman's (2002) investigation of knowledge of derivational forms can be seen as having some relevance to the present study in that it investigates indirectly the use by novice writers of inflected and derived forms of the words they know through an analysis of their use of MWUs such as *to decide* versus *make a decision*. In their study, Schmitt and Zimmerman (2002:163) suggest that certain learners may 'plateau' at a particular level in their learning of derivational forms, partly because they can communicate quite easily without a working knowledge of all the derivative forms in a word family. This means that they lack what Laufer (1991:441) has called the 'communicative need' which might drive them to learn additional words. This may relate particularly to derivational forms where learners can make themselves understood quite adequately when using the wrong form of a particular word. The students in Schmitt and Zimmerman's (2002) study might be examples that support Laufer's 'active vocabulary threshold hypothesis', in that their passive vocabulary continues to develop throughout their lives, but their productive vocabulary will grow 'only until it reaches the average level of the group in which [they] are required to function' (Laufer 1991:445).

Besides the number of words known, the present study also investigates, albeit in some cases indirectly, aspects of collocation, polysemy and synonymy, given that it analyses novice writers' production of MWUs featuring delexical, high-frequency verbs which are regarded as being highly polysemous. As noted above (§1.3.2) scholars such as Viberg (1996, cited in Altenberg and Granger 2001) and Altenberg and Granger (2001) have referred to delexical verbs as polysemous, as do Biber et al. (1999:412) (see §5.9.3.3). In this sense it contributes to the body of research on depth of vocabulary knowledge.

In the following sections the focus moves from depth and breadth of vocabulary knowledge to the issue of vocabulary lists and the debates surrounding the use of these in teaching and learning.

3.4 WORD LISTS AND ACADEMIC VOCABULARY

Electronic corpora have brought new life to vocabulary studies, and corpus studies have shown that West's (1953) General Service List (GSL) (see §3.2.1) of basic vocabulary is still very useful (Hancioğlu et al. 2008:460). Since its development in the 1950s, this list has been followed by further specialised lists of vocabulary such as Xue and Nations's (1984) University Word List (UWL) and Coxhead's Academic Word List (AWL) (Coxhead 2000).

The question of what and how many words a learner needs to know to succeed in academic study continues to be debated. In an effort to establish more clearly what words language learners in an L2 environment need in order to cope with academic studies, Hancioğlu et al. (2008) examined the GSL and the AWL, not only because these lists are very commonly used, but also 'because for ESP practitioners they offer a package more or less suggesting that the basis for survival in an academic environment is knowledge of the 2000 word families of the GSL plus the 570 word families of the AWL' (2008:461). One of the aims of their study (2008:462) was to refine the GSL, creating a stronger and more relevant list. They also examined the AWL (Coxhead 2000) which has come in for a fair amount of criticism over the years as attitudes to and understanding of academic vocabulary have developed. Hancioğlu et al. (2008) took a 'corpus-informed approach' to a review and revision of the GSL, a reconceptualisation of the AWL by integrating it into the revised GSL, and the creation of a model for 'lexico-structural banks' that could be used in teaching, particularly the teaching of thesis writing to postgraduate NNS language learners (2008:465).

In one aspect of the study reported on in Hancioğlu et al. (2008:469ff), two corpora comprising abstracts from theses were compiled, the target abstract corpus (TAC) and the learner abstract corpus (LAC). The LAC abstracts came from the disciplines in which postgraduate students in an advanced thesis writing course were enrolled: Arts and Humanities, Sciences, Social Sciences and Architecture. The abstracts in the TAC were taken from universities in countries where English was the native language, but no distinction was made between native and non-native speakers of English. A list of the 165 most frequent sub-technical content words (termed 'keywords') was generated from this corpus of learner writing, and it was found that they came from both the AWL (85 word families, including *analyse*, *design*, *process*) and the GSL (59 word families, including *apply*, *consider*, *aim*, *combine*), with 21 words unaccounted for by either list (such as *dissertation*, *interview*, *collaborate*) (Hancioğlu et al. 2008:468); the GSL contained some words which were common in academic usage, while the AWL in its turn featured words which were frequently found outside academic texts (Hancioğlu et al. 2008:470–1). These words are what Hancioğlu et al. (2008:465)

would call 'scaffolding' words and what scholars such as Paquot (2010) and Schmitt et al. (2001) would call sub-technical words, but Hancioğlu et al. (2008:471) observe that it was the 'co-occurrence' of these words 'in the same environment' that 'gave the list its academic texture'. It was concluded that there was no real reason for making a division between the GSL and the AWL (Hancioğlu et al. 2008:471). The authors believe that these findings

> confirmed the initial hypothesis that teaching academic writing needed to be based on analyses of how lexical items collated and combined to achieve specified moves for specified purposes within particular genres, and that many lexico-structural building blocks that were emerging were cross-disciplinary rather than disciplinary specific (Hancioğlu et al. 2008:471).

This remark poses some opposition to Hyland and Tse's (2007) argument for the value of genre-based lists of academic vocabulary. These scholars criticise Coxhead's approach to data collection for the AWL, arguing that her list is biased in favour of certain academic fields. They dispute the accepted belief that there is 'a single literacy which university students need to acquire to participate in academic environments', that is, the existence of a single category of academic vocabulary, and that one list of academic words will suit all students, regardless of their discipline (Hyland and Tse 2007:236). They found that the AWL covered 'an impressive 10.6% of the words' in their corpus, and together with the 2000 words of the GSL it provided an accumulative [sic] coverage of 85%, meaning that readers would encounter about one unknown word in every seven words of text' (2007:240). They conceded that this was good overall coverage but found that there was not an even distribution: the combined AWL and GSL failed to account for 22% of the words in their science corpus; students with this level of knowledge would thus encounter an unknown item on average once in every five words, which might make the text incomprehensible. Hyland and Tse (2007) argue thus for lists for specific domains or disciplines which contain more specialised and technical vocabulary, or for a general core of academic vocabulary together with specialised lists which are discipline and genre specific. They believe that all disciplines 'shape words for their own uses' (2007:240) and this hampers any attempts to create a list of core academic vocabulary. They conclude that

> a perspective which seeks to identify and teach such a vocabulary [that is, a general list of academic vocabulary] fails to engage with current conceptions of literacy and EAP, ignores important differences in the collocational and semantic behavior of words, and does not correspond with the ways language is actually used in academic writing (Hyland and Tse 2007:236-7).

Although this is supported in the claim by Hancioğlu et al. (2008:471) that the teaching of academic writing should be based on 'analysis of how lexical items collated and combined to achieve specified moves for specified purposes within particular genres', the latter's further claim that 'many lexico-structural building blocks that were emerging were cross-disciplinary rather than disciplinary specific' (Hancioğlu et al.

2008:471) would seem to dispute it. But despite this lack of agreement in the approach to and content of academic word lists, Hyland and Tse's (2007) findings do support the ever-increasing awareness of the importance of studying not just words as discrete, isolated entities, but rather how these words behave in context, in a corpus.

Paquot (2010) has also joined the debate on the usefulness of established, itemised wordlists in the teaching and learning of vocabulary in the development of her Academic Keyword List (AKL), a 'productive counterpart to the AWL (Coxhead 2000)' (2010:212). She questions the 'well-established frequency-based distinction between general service words and academic words' (Paquot (2010:212), and instead includes in the procedure used to identify words for her AKL the first 2000 core vocabulary words as well as academic words. She believes that we 'should not assume that EAP students [or ESL, for that matter] know the first 2000 words of English' – these words are very important, especially to academic writing, as they have 'discourse-organizing functions' (2010:212).

In compiling her AKL, Paquot extracted keywords from corpora of both professional and student academic writing 'corresponding to five broad academic domains', such as arts and social science (Paquot 2010:31). Those words she selected for her AKL met the definition of academic vocabulary: 'a set of options to refer to those activities that characterize academic work, organize scientific discourse and build the rhetoric of academic texts' (Paquot 2010:28), and were refined further by the criteria of range and evenness of distribution through the corpora. This resulted in a list of 599 potential academic words, including nouns such as *conclusion*, *extent*, *significance*, verbs such as *prove*, *appear*, *discuss*, adjectives such as *significant*, *effective*, *similar*, and adverbs such as *particularly*, *conversely*, *highly* (Paquot 2010:52).

In Paquot's (2010) view these words should in fact be the focus of teaching, especially teaching which is aimed at writing activities. She also believes that there needs to be a clear distinction between the vocabulary which is needed for academic reading and that which academic writing requires. This difference may be quantitative rather than qualitative: perhaps it is not a case of different vocabulary, but rather the quantity of words recognised and known in depth; students may be able to get by with fewer words in their writing than in their reading (see §3.2.1).

The debate on the usefulness of established word lists, and the compilation of new ones, has also led to increased questioning of the value of lists of single words on their own. Of particular relevance to the present study, Hyland and Tse (2007:246) found that particular 'lexical bundles', or strings of words that 'commonly go together in natural discourse', also contribute to meaning-making in academic contexts. These fixed or semi-restricted word combinations are an 'important part of a discipline's discoursal resources but enormously complicate the business of constructing general word lists' (Biber, Johansson,

Leech, Conrad and Finegan 1999:990). However, breaking these into single word items which can be more easily learnt as wholes misrepresents discipline-specific meanings and misleads students (Hyland and Tse 2007:241).

In response to this increasing awareness that 'language is made up of not only individual words, but also a great deal of formulaic language' (Martinez and Schmitt 2012:299), Shin and Nation (2008), Simpson-Vlach and Ellis (2010) and Martinez and Schmitt (2012) have all compiled lists of MWUs which they believe should form part of teaching materials and practice. Citing researchers such as Pawley and Syder (1983) in their assertion that collocations are useful to learners both in developing fluency and in choosing the sort of collocations native speakers would use, Shin and Nation (2008:340) set out to establish criteria to distinguish collocations from other MWUs and to identify the most frequent collocations and the commonest collocation patterns in order to derive a list of spoken collocations that would be most useful for elementary learners of English. According to their criteria, a collocation is 'a group of two or more words that occur frequently together, and it is not restricted to two or three word sequences' (2008:341). Each collocation has a 'pivot' word, or a focal word, and Shin and Nation (2008) used these when searching for collocations. Using the spoken section of the BNC and WordSmith Tools, they applied six criteria to identify the collocations they wished to extract. These criteria included the frequency of these collocations – each pivot word had to occur within the most frequent 1000 words of English, and each collocation had to occur at least 30 times per 10 million words (2008:341–342).

Their most striking finding was the large number of collocations that met the six criteria, and the fact that a large number of these would qualify for inclusion in the most frequent 2000 words of English if no distinction were made between single words and collocations. Moreover, many of these collocations could be usefully taught in an elementary speaking course. They found that there were many 'grammatically well-formed high frequency collocations' (Shin and Nation 2008:343) in their corpus, and the higher the frequency of the pivot word, the greater the number of collocates it formed. Conversely, only a few pivot words accounted for a large number of the tokens of collocations. Lastly, the shorter the collocation, the greater its frequency (2008:344). Comparing their findings to those of a study conducted by Shin (2007, cited in Shin and Nation 2008:344), they found that collocations were far more frequent in spoken than in written language, underlining the importance of including these MWUs in language teaching courses focusing on spoken language.

Simpson-Vlach and Ellis (2010:487) compiled a list which they called the Academic Formulas List (AFL), comprising formulaic sequences which occur frequently in academic speech and writing. This list was intended to be comparable to the AWL. In their study, they combined quantitative and qualitative criteria, corpus statistics, psycholinguistic processing metrics and insights from teachers to establish a measure of

utility called 'formula teaching worth' (2010:488) to rank the formulas on this list. Although their starting point in choosing the sequences for their list was frequency of occurrence, they believe that their approach was 'substantially more robust' (2010:490) than earlier corpus-based methods in that it included a mutual information (MI) score, a statistical measure of collocational strength between items. They used the MICASE corpus[14] together with the BNC files of academic speech, and Hyland's (2004, cited in Simpson-Vlach and Ellis 2010:490) research article corpus and selected BNC files of academic writing from which to extract their data. They also used the Switchboard (2006, cited in Simpson-Vlach and Ellis 2010:490) corpus of non-academic speech and the FLOB and Frown corpora of non-academic writing for comparison purposes (ICAME 2006, cited in Simpson-Vlach and Ellis 2010:490). They included in their data strings of three, four and five words, and used a lower cut-off range of 10 occurrences per million words. Once overlapping data had been removed, they were left with a list of about 14 000 sequences. In order to 'sift out' (2010:492) those sequences which occurred most frequently in both academic and non-academic discourse, so that they could extract those that were characteristic of academic writing, Simpson-Vlach and Ellis (2010) used the log-likelihood (LL) statistic to compare the relative frequency of the sequences across registers.

The final AFL comprises those sequences which were found to occur most frequently in academic discourse. Once items that occurred in both academic speech and academic writing had been identified, the distribution of these words across the academic subdivisions of the corpora was established: the criterion for their inclusion was an occurrence of at least 10 times per million words in 'four out of five of the academic divisions' for expressions in speech, that is Humanities and Arts, Social Sciences, Biological Sciences, Physical Sciences and 'non-departmental or other' (Simpson-Vlach and Ellis 2010:491), resulting in a Spoken AFL of 979 items, and at least 10 times per million words in '*three out of four* academic divisions', that is Humanities and Arts, Social Sciences, Natural Sciences/Medicine and Technology and Engineering. for expressions which occurred mainly in writing, making a Written AFL of 712 items (Simpson-Vlach and Ellis 2010:493). Items which occurred in both speech and writing had to occur at least 10 times per million in at least six of the corpora, resulting in a list of 207 Core AFL items. Finally, these researchers used the 'formula teaching worth' (FTW) (Simpson-Vlach and Ellis 2010:488) score which indicates 'a measure of utility that is both educationally valid and operationalizable with corpus linguistic metrics' (2010:508) to identify those formulas which would be most useful for English for academic purposes (EAP) teaching. This score compares MI and frequency.

[14] The Michigan Corpus of Academic Spoken English (MICASE), a collection of nearly 1.8 million words of transcribed speech from the University of Michigan (U-M) in Ann Arbor. Original compilers: Rita Simpson-Vlach (project manager 1997 to 2006); John Swales (faculty advisor); Sarah Briggs (testing advisor). Available online at http://quod.lib.umich.edu/m/micase/ [Accessed 7 March 2014]

Simpson-Vlach and Ellis (2010:508) found that items which achieved the highest scores on the FTW measure did indeed appear to be 'more formulaic, coherent, and perceptually salient'. The authors thus used the FTW scores as the basis for the selection of items for the Core and top 200 Written and Spoken formulas. One important point which emerged clearly from this study was that there are certainly enough lexical bundles that occur across multiple disciplines and from which a core list of formulas can be compiled, contrary to what Hyland (2008) claims. Simpson-Vlach and Ellis succeeded in compiling a list of common core academic formulas which transcends the boundaries of academic disciplines.

Subsequent to this research study by Simpson-Vlach and Ellis (2010), Martinez and Schmitt (2012) compiled their PHRASal Expressions (PHRASE) List. They, too, argue that such formulaic language should form a part of language teaching textbooks and practice (2012:301). They make the point that many formulaic sequences are difficult for learners to understand, even when they appear simple to native speakers, and like Simpson-Vlach and Ellis (2010), they wished thus to create a list which would 'have pedagogical utility' (2012:302). Informed by research by Nation (2006), who found that 6000 to 7000 word families were required to comprehend a range of spoken discourse, and 8000 to 9000 for written discourse, Martinez and Schmitt compromised between these requirements and those of a list of the size of 2000 words like the GSL and decided to include phrases which were of the same frequency as words up to the 5000-word level, which represents, as Read (2000, cited in Martinez and Schmitt 2012:303) notes, 'the upper limit of general high-frequency vocabulary'. In order to be of greatest use, Martinez and Schmitt (2012:303) included only those formulaic sequences 'that realize meanings or functions' and used this as one of their selection criteria. They also included only those phrases that were 'uncompositional', that is, where the meaning could not easily be derived from the meaning of the individual words. Their chief criteria were thus frequency, meaningfulness and relative non-compositionality (2012:304). In naming their list, Martinez and Schmitt (2012:304) chose the term 'phrasal expressions', which they defined as

> a fixed or semi-fixed sequence of two or more co-occurring but not necessarily contiguous words with a cohesive meaning or function that is not easily discernible by decoding the individual words alone.

Once they had decided on their criteria for selection, Martinez and Schmitt (2012:310) selected the BNC as their data source as it represents both written and spoken language and, despite minor limitations, is large and has a history of use in research. Using a mixed-method corpus analysis, they compiled a final PHRASE List of 505 multiword items which, if treated as part of the 5000 most frequent word families would make up over 10% of the total items on the list (2012:313). This they believe is evidence of the pervasiveness of formulaic language, and counters assertions by researchers such as Moon (1998a:63) and Grant and Nation (2006, cited in Martinez and Schmitt 2012:313) that 'the number of commonly occurring opaque multiword expressions in English is low' (Martinez and Schmitt 2012:313). Martinez and Schmitt (2012:313) also found that these 505 phrasal expressions were made up almost entirely of words from the top 2000 words of

English, the majority of which were from the 1000-word level, bolstering arguments that these high-frequency words are 'merely the tips of phraseological icebergs', and providing further support for the investigation of high-frequency verbs in the present study. Martinez and Schmitt (2012:313) also note that, when these phrasal expressions (or MWUs) are taken into account in the analysis of an academic text, the number of words which do not occur on any list rises from a 'relatively manageable' 7.46% to a 'much more onerous 26.87%'. Readers who do not know these expressions will thus have a considerably lower percentage of text coverage.

As far as academic vocabulary is concerned, scholars have disputed how much students should know. Hyland and Tse (2007) have challenged the use of decontextualised lists of words, supposedly equally valid items, for student writers across the disciplines. They believe that although some words appear to be more generally used in academic texts, ranging across several disciplines, these items are not used in the same way and do not mean exactly the same thing in different disciplinary contexts. As Biber et al. (1999) note, words are often associated with different meanings and uses across registers and there are similar variations across different disciplines. Hyland and Tse (2007:249) suggest instead that we regard academic vocabulary as 'a cline of technically loaded or specialized words ranging from terms which are only used in a particular discipline to those which share some features of meaning and use with words in other fields'. They believe that we should identify and address students' target language needs by making sure they are introduced to and have opportunities to use the specialised vocabulary in their disciplines (Hyland and Tse 2007:249). This supports the point which Hyland and Tse (2007) make about the importance of context, and the implication that the words with which a word keeps company often affect its meaning.

Taking Hyland and Tse's (2007) notion of a cline of academic vocabulary a step further, Hancioğlu et al. (2008:465) warn against 'discarding' the AWL in favour of more specialised lists which, they believe, may have adverse effects on students' writing by doing away with the teaching of words which play a scaffolding role in English. They believe that what students really need are those more general words which are used for writing and writing about academic tasks (they cite McCarthy and O'Dell (2008, cited in Hancioğlu et al. 2008:463) in this regard). Paquot (2010) believes too that learners need both high-frequency words as well as core academic vocabulary that occurs across disciplines. Hunston (2002:135) agrees that it is what she calls 'the terminology of rhetoric' which learners need to allow them to cope with the demands of academic writing, rather than the technical words specific to their field. She believes that this is particularly relevant to academic experts who write papers in languages other than their own, but could apply equally to students such as the novice writers in this study.

The jury is still out on the usefulness of single-word lists such as the AWL and the GSL, but these continue to be used in research and to provide reliable results. For instance, the present study could provide further

insight into Hyland and Tse's (2007:249) claim that there is 'a cline of technically loaded or specialized words', given that students' scores on the discrete test of the UWL (Level 4 in the test used in my study) could provide some indication of where on this cline students are, and how this relates to their production of MWUs and academic performance. This study also hopes to add new information on students' knowledge and use of MWUs in academic writing, the focus of much recent research. What is clear, though, is that despite the debates on what vocabulary should be presented to learners and how, it remains a vital aspect in the mastery of a language, and electronic corpora are changing the ways this is investigated and taught. Hancioğlu et al. (2008:461) sum it up thus: 'wordlists and corpora unquestionably offer a portal into the complex behaviour and intricate relationships of individual lexical items'.

The following section addresses what research has revealed about the effects of an inadequate vocabulary, both at school and at university.

3.5 THE RELATIONSHIP BETWEEN VOCABULARY AND ACADEMIC PERFORMANCE

This section attempts to throw some light on the importance of vocabulary, both academic and high-frequency words, for learners in the formal learning environment. By the time they reach university, students should have not only a good receptive command of the vocabulary of English, in order to read and understand the literature in their discipline, but also a productive command of the language to allow them to master the writing of academic essays, articles, dissertations and theses (Paquot 2010:1). Paquot (2010), Reynolds (2005), Nation and Waring (1997) and Evans and Green (2007) all stress the importance of an 'advanced linguistic competence' for academic writing, which demands a 'range of lexical and grammatical skills' (Paquot 2010:1).

3.5.1 The role of vocabulary at school

In the light of the above, this section focuses on what is occurring in schools, as this is where the foundation of a vocabulary that allows a student to read with comprehension should be laid. McCormick (1995) describes reading performance according to four levels: the independent level (at which learners are able to read on their own, recognising, that is, being able to decode[15], 99% of the words in context and enjoying 90% comprehension); the instructional level (at which readers can profit from instruction, recognising 95% of words in context, and enjoying 75% comprehension); borderline (90–94% decoding accuracy and 55–74% comprehension accuracy); and the frustration level (at which readers are unable to cope with the reading matter, recognising fewer than 90% of words in context, and comprehending about

[15] Decoding refers to automatic word recognition, a phonological-graphemic process, and not lexical access. This means that a student may be able to decode a word, without necessarily understanding what it means.

50% or less of the meaning of the text). Reading comprehension feeds naturally into the ability to write competently, but without enough words, learners are unlikely to read independently or to build their vocabulary in this way. Nation and Waring (1997:11), referring to beginning learners and readers, observe that 'if one does not know enough of the words on a page and have comprehension of what is being read, one cannot learn [words] from context'. This will necessarily affect the quality of learners' writing. In his study, Reynolds (2005:27) investigated the development of linguistic fluency in the writing of American fifth to eighth grade school children, some enrolled in an ESL class and others in a 'regular language arts' or RLA course (many of whom indicated Spanish as a second language spoken at home). He found, in support of previous studies, that the second language learners needed to improve both the 'size and sophistication of their lexicon' (2005:41): the writing of the ESL group lacked the informational density of the RLA group, and as far as narrative was concerned, the RLA students' writing was noticeably more diverse than that of the ESL students.

In a study of South African Grade 7 ESL learners[16] in an immersion situation, Scheepers (2003) found, contrary to expectations, that although there were significant differences in receptive vocabulary scores (as measured by the VLT) between late and early immersion students, and between immersion students and those who were NSs of English, these differences (in favour of early immersion and NSs) were not reflected in these learners' productive vocabulary. In a free productive writing exercise very few of the children, from any of the three groups, early immersion, late immersion or NSs of English, used vocabulary from levels more advanced than the 2000-word level. Whatever the reasons for this – including possibly the fact that the writing test was a free productive exercise and that the participants wrote very little on the whole – all the children in this particular study, regardless of their language background, revealed a limited productive vocabulary, comprising predominantly words from the 1000- and 2000-word levels, that is, basic high-frequency vocabulary. This suggests that the classroom activities in which these learners were engaged did not require extensive reading, which would have exposed them to words beyond the 2000-word level, nor were writing activities in the classroom sufficiently demanding. This was particularly concerning as these children were in their final year of primary school; they would soon be entering high school, where the demands on their reading and writing skills would increase significantly.

The PIRLS 2011 study (Howie et al. 2012) found that, as in PIRLS 2006, the quality of reading instruction in South African schools still requires interrogation. Both prePIRLS and PIRLS 2011 revealed that South African children at Grade 4 and 5 level, even those at the best schools, performed consistently lower in reading literacy than the International Centre point score of 500. This has implications for vocabulary development: if learners are reading so poorly, they will not develop their vocabulary to a level at which they will be able to read independently. In Scheepers's (2003) study, these low reading expectations were reflected in the

[16] 'Learners' here is used in the sense of 'school pupils', as required by the South African Department of Basic Education.

number of words children in Grade 4 were expected to know: 1000 to 2500 words. Since that study was conducted, the vocabulary requirements for Grade 4 First Additional Language have been set at a more realistic level of 2000 to 3500 words (Curriculum and Assessment Policy Statement (CAPS) n.d.:27), with learners expected to know 6000 words by the time they reach Grade 7.

These findings lend support to Paquot's (2010) belief that the difficulties experienced in academic writing cut across the board; both novice and advanced L2 learners and novice L1 writers experience them (Paquot 2010:2). Hinkel (2002, cited in Paquot 2010:3) believes, too, that the difficulties begin at school level as NNS students are unprepared for academic writing when they arrive at higher education institutions; he reported that the focus in the classroom is on a process writing approach, with little direct teaching of grammar and the frequent neglect of academic vocabulary. Sun and Feng (2009) cite Stanley (in Sun and Feng 2009:150) in their explanation of process writing as treating all writing as 'a creative act which requires time and feedback to be done well'. Such an approach focuses on the steps involved in the creation of a written text, while the more traditional, product writing approach focuses on producing texts that are coherent and free of errors (Sun and Feng 2009:151)

That vocabulary knowledge is vital to reading has thus long been accepted. Laufer (1986:69) believes that '[n]o language acquisition can take place without the acquisition of lexis' and '[r]eading comprehension is strongly related to vocabulary knowledge, more strongly than other components of reading' (Laufer 1997:20). Also, vocabulary size has proved to be a good predictor of reading success in second language studies (Cooper 1999; Laufer 1992; Qian 2002). Coady et al. (1993) showed that greater proficiency in high-frequency vocabulary also led to improved reading proficiency. Milton and Treffers-Daller (2013) cite others (Biemiller 2001, cited in Milton and Treffers-Daller 2013; Hart and Risley, 2003, cited in Milton and Treffers-Daller 2013) who support the claim that vocabulary size can have significant implications for subsequent school attainment. Morris and Cobb (2004:79) note that 'the more exposure people have to words, the more they acquire. The more time spent on grammatical explanations, the more proficient they become'. But the reverse may also be true: if children are not given the tools to develop into independent readers, they will be unable to build their vocabulary and their knowledge through reading. This could be particularly pertinent to the sample in the present study, many of whom may have come from home and school backgrounds which are print poor, and where reading is not regarded as a priority. Also, their models in the classroom may not have been NS or near-NS in proficiency. Implicit in this is the value of extensive reading for success in academic pursuits, both reading and writing. This is an aspect which is often neglected in South African schools, as has been highlighted by studies such as PIRLS 2006 and 2011 (Howie et al. 2012; Howie 2010). This is not directly addressed in this study but it is a reality which certainly underlies students' proficiency in English.

The following section focuses on the importance of vocabulary for university students, investigating in greater depth the link between vocabulary knowledge and academic success, which is the particular focus of Phase 3 of this study.

3.5.2 Vocabulary knowledge and success at university

Research has shown a relationship between vocabulary knowledge (more with regard to breadth, though occasionally also regarding depth) and success at university, that is, academic performance. For instance, Cooper (1999, 2000), in an examination of the writing by first level ESL university students at a South African institution, found that there was a relationship between the breadth of the vocabulary knowledge these students had and their academic performance: weaker students had small receptive vocabularies and lacked knowledge of lower frequency (academic) words. These students mostly had neither the high-frequency word knowledge nor the knowledge of academic vocabulary (as reflected in the University Word List [UWL]) to cope with academic texts (Cooper 1999; 2000:28). Cooper's analysis of first-year course material for the development of vocabulary tests in her study revealed that academic vocabulary made up 9.7% of the running words in a text, and 20.3% of the total number of lexemes. She illustrated her findings on the relative proportion of academic vocabulary to basic (the 1000 and 2000 word level) and advanced vocabulary (the latter is defined as 'low frequency, narrow range vocabulary items which occur only as technical terms in a specialised field, for example, *cataphora*, *phoneme*, *sintisoidal* and *thixotropy*' (Cooper 2000:20) as follows:

Frequency list	Percentage of tokens	Percentage of types
1000	77.1	37.0
2000	6.5	13.3
UWL	9.7	20.3
Other	6.7	29,4

(Cooper 2000:21)

Table 3.1: Proportion of academic to basic and advanced vocabulary

From this breakdown, Cooper concluded that high-frequency words (the 1000- and 2000-word levels), what she terms basic vocabulary, make up 83.6% of the running words of a text and half the lexemes. The tests of vocabulary size which Cooper (1999) used revealed that L2 students' 'overall grasp' of academic vocabulary was not adequate in meeting the lexical demands of their reading at university level (Cooper 2000:19). She found that almost half (45%) of the students who failed the academic vocabulary test failed the year (Cooper 1999:88). Her study supports the notion that receptive vocabulary must reach a level, or 'threshold', before learners can transfer their L1 reading skills to their L2 (Cooper 2000:19; Nation and Waring 1997). Nation (1993:131), too, believes that 'vocabulary size is the essential prerequisite for the development of skill in language use'. As it grows, this skill allows for a growth in knowledge of the world [academic ability] through the competent use of the language. If this knowledge is to increase, vocabulary

must also increase. Thus, skill in language use, which includes reading comprehension, is vital to success at university and is dependent on vocabulary size (Nation 1993:120).

Regarding the number of words needed by students to succeed in university studies, Schmitt et al. (2001:56) believe that L2 learners with a knowledge of the most frequent 10 000 words in English can be considered to have a wide vocabulary, and agree that a learner may need a vocabulary of this size to cope with university study in a second language; however, such learners will also need knowledge of academic vocabulary (what Schmitt et al. (2001:56) and Paquot (2010) refer to as 'the sub-technical vocabulary that occurs across a range of academic disciplines' and what Hyland and Tse (2007) call a 'common core of academic vocabulary'). This underlines the caution with which we should view estimates of the size of vocabulary required by learners at university, voiced by both Vermeer (2001) and Milton and Treffers-Daller (2013); such estimates range from as few as 9000 words to as many as 155 000 words.

More recently, Milton and Treffers-Daller (2011, 2013:165) have queried Schmitt et al.'s (2001:56) view that L2 learners with a knowledge of the most frequent 10 000 words in English have a wide vocabulary. Milton and Treffers-Daller (2011, 2013) note that Nation (2006) suggests that 8000 to 9000 words are necessary for general reading of newspapers and novels; Nation uses a figure of 98% coverage as the basis for this estimate. On the other hand, in their study of Dutch-speaking and English NS students, Hazenberg and Hulstijn (1996:158) found that '[i]ndividuals with a vocabulary of fewer than ten thousand base words run a serious risk of not attaining the reading comprehension level required for entering university studies'. With these estimates in mind, Milton and Treffers-Daller (2011, 2013) investigated how many words a group of first-, second- and third-year students between 18 and 19 years of age at three universities in South West England needed to read their textbooks. They were also interested in investigating any relationship between the number of words their students knew, their reading habits, and their academic achievement. In order to test students' vocabulary knowledge, they designed two tests, one made up of a 50-word sample from Thorndike and Lorge's (1944, cited in Milton and Treffers-Daller 2013:160) frequency lists and the other comprising 221 words which fell outside Thorndike and Lorge's 25 000 word range, sampled from Webster's Dictionary. In addition, students were asked to complete a questionnaire providing information on how many books per year and newspapers per day they were in the habit of reading.

Milton and Treffers-Daller (2013) found that, on entering university, these students had smaller vocabularies than had been suggested by earlier studies; they knew about 10 000 words, and about 11 000 in their final year, a much smaller number than earlier studies they cite, such as Nagy and Hermann (1987, cited in Milton and Treffers-Daller 2013:165). The students in their study represented three language groups: non-native English, bilingual and monolingual English. The non-natives' vocabulary was statistically

different from the other two groups; the difference between the native and bilingual groups was too small to be statistically significant. As far as the link between vocabulary size and reading habits was concerned, there was no correlation but the authors concede that this may have been because 'our measure of reading habits was rather crude' (Milton and Treffers-Daller 2013:167–168).

As far as the present study is concerned, Milton and Treffers-Daller's (2013) finding that there were no significant correlations between vocabulary size and reading is interesting as it may suggest that literacy levels are dropping, even in countries where education systems are for the most part functional. Are undergraduates reading less, and are they also required to read less for their degrees? These questions are certainly relevant to the situation in South Africa, where literacy levels in schools are very low (as discussed in §1.2) and where students at university also exhibit difficulties in reading and writing stemming from a lack of vocabulary (Cooper 2000:29). The fact, too, that Milton and Treffers-Daller (2011, 2013) found that native speaker vocabulary size appeared much smaller among their undergraduates than had been previously assumed and that scores were comparable in scale to able L2 learners (Schmitt 2008) may also be an indication of a drop in reading practice. However, it is possible, of course, that these findings are the result of their methods. Questionnaires can be notoriously unreliable and social desirability effects in reading questionnaires may skew relationships; in other words, respondents may provide the type of answers that they perceive as desirable in the circumstances, rather than their true feelings or opinions. As noted above (§3.3.1), Lukhele (2010) and Oyetunji (2011) both found discrepancies between their respondents' self-evaluations and their performance on reading and vocabulary tests.

Qian (2002:517–518) (see §3.3.1) argues that both vocabulary size and depth dimensions are important in reading comprehension. He believes, as do many others (Hazenberg and Hulstijn 1996; Laufer 1991; Milton and Treffers-Daller 2011, 2013; Morris and Cobb 2004; Schmitt and Zimmerman 2002; Vermeer 2001), that vocabulary is acquired in an incremental fashion, with those words which are acquired at the beginning of the learning process likely to be learnt in more depth than those learnt later in the process; and the more words a learner knows, the deeper the knowledge of all these words is likely to be. Depth of vocabulary knowledge is also likely to improve learners' ability to guess the meaning of unfamiliar words in context, and so the two dimensions of lexical knowledge support each other (Qian 2002:518).

In a study that relates closely to that of Qian (2002) and Vermeer (2001), Akbarian (2010) investigated whether there was a relationship between vocabulary size and vocabulary depth among Iranian postgraduates enrolled in an English for Academic Purposes (EAP) course. Using a passive version of the VLT to measure size, and Read's (1993, cited in Akbarian 2010) Word Associates Test (WAT) to measure aspects of depth of vocabulary knowledge such as synonymy and collocation, he found that the two tests correlated strongly when the results for the group as a whole were calculated. In order to achieve a clearer

idea of whether size and depth correlated across levels of proficiency, Akbarian also reported the results of the high and low performers in his group. This division was based on whether students had achieved mastery of the 2000-word level. Only a small portion of the group achieved this and made up the 'high proficient' group while the remainder made up the 'low proficient' group (Akbarian 2010:399).

Using linear regression analysis, Akbarian found that vocabulary size was a strong predictor of vocabulary depth for the high proficient group. As far as the low proficient group was concerned, although there was a strong correlation between the VLT and the WAT, vocabulary size was not as strongly predictive of vocabulary depth as it had been in the stronger group. In other words, the more words students knew, the greater the depth of their vocabulary knowledge was likely to be. Akbarian (2010) found that his students as a group lacked both size and depth of vocabulary knowledge; weaker students lagged further behind than the stronger group in vocabulary size, but even in the stronger group, only two students had mastered the 3000-word level. He observes that this 'low vocabulary proficiency level' of all the ESP/EAP learners at his institution was a cause for grave concern (Akbarian 2010:399).

Akbarian (2010) posits the notion that learners such as those in his study first learn a number of words and then begin to build up a network of knowledge about these words – but this can only occur once their vocabulary size has reached a certain level. He concludes that 'while breadth and depth of vocabulary knowledge might converge when language learners are relatively advanced, the dimensions are more distinct at lower levels of language proficiency' (Akbarian 2010:400).

Morris and Cobb (2004:78) mention several studies (Cobb and Horst 2001, Coxhead and Nation 2001, cited in Morris and Cobb 2004; Laufer and Nation 1995; Laufer and Paribakht 1998) that have indicated the 'key role' played by the knowledge of academic vocabulary in academic success among second-language learners. Many of these words 'play a metalinguistic or metacognitive' role in discourse (Morris and Cobb 2004:79), which is vital to success in academic studies. In a study to test the viability of vocabulary size as a predictor of academic success, Morris and Cobb (2004) used Vocabprofiler (Cobb n.d.) to analyse the vocabulary and to predict the linguistic and academic ability of bilingual or multilingual Canadian TESL (Teaching English as a Second Language) trainees in a BEd programme. These students all had native-like or near native-like proficiency in spoken English and high levels of proficiency in written English and, as in the case of the students in my study, they came from a variety of L1 backgrounds.

In order to establish the vocabulary profiles of their students, Morris and Cobb (2004) used a corpus comprising their students' entrance exam essays. The first 300 words of each essay were then entered into Vocabprofiler (Cobb n.d.) and the type-token ratio (TTR) and percentage of words in various groups were calculated. In order to measure academic success, the researchers used the marks achieved by students on

two compulsory grammar courses, which they called G1 (an introduction to English grammar focusing on *knowing that,* or declarative knowledge) and G2, the second pedagogical grammar course focusing on *knowing how* (Morris and Cobb 2004:81). They then measured the correlations between the various frequency ranges of the vocabulary profiles and the different grammar course results.

Morris and Cobb (2004) found that the highest correlation (r = 0.37) was between academic words (the AWL) and grades in G2, or procedural knowledge. Two other significant correlations were found between the first thousand words and the G2 scores, and between function words and the G1 course scores (r = 0.34 in both cases). Although these correlations were not high enough to justify using a vocabulary profile as a 'stand-alone instrument' (2004:82) for TESL candidate assessment, it was particularly interesting to these researchers that the correlations occurred with grades in the course which dealt with teaching grammar. This is a reminder of Hinkel's (2002, cited in Paquot 2010:3) view that grammar and vocabulary teaching are neglected in schools to the detriment of students' writing skills at university (see §3.5.1).

From a correlation of all the vocabulary profile scores of these students with their academic grades, Morris and Cobb (2004) established the percentage of coverage of students' texts which resulted in academic success: a score on the first thousand words of at least 85%, a score on the AWL of over 5%, and a score on function words of over 50%. The NSs outperformed the NNSs on every comparison, achieving higher marks and using fewer words from the first thousand level and function words and more from the AWL (Morris and Cobb 2004:82–3). They found that, according to this standard, NNSs were 50% less likely to do as well as NSs – 'the lexical playing field was not a level one for NS and NNSs, even though it would have appeared to have been had the analysis been limited to interview assessments of oral proficiency' (2004:83). Less than half the group of NNSs (46%) achieved the ideal profile on the AWL, suggesting that academic words present a particular stumbling block to ESL learners. It is also significant that the authors believe that the profile would have been very different had they confined their data to oral interviews, as BICS, or basic interpersonal communicative skills, require a different type of vocabulary, distinct from CALP (cognitive academic language proficiency), that is, the ability to read and write in an academic register. This relates to Laufer (1991) and Schmitt and Zimmerman's (2002) remarks about L2 learners reaching a plateau in their language, a plateau which sometimes only becomes noticeable in their writing. Morris and Cobb's (2004) study adds further weight to the widely held belief that academic vocabulary is vital to success in university studies.

In a study of Cantonese-speaking students at an English-medium university in Hong Kong, Evans and Green (2007) found that students had difficulties with academic reading, particularly with understanding subject-specific (or 'technical') vocabulary as well as 'difficult words', those sub-technical or 'common core' academic lexical items found in most disciplines (Evans and Green 2007:11; Paquot 2010). Like Hyland and

Tse (2007) and Paquot (2010), Evans and Green (2007:14) concluded that EAP programme design should place more emphasis on the teaching and learning of both subject-specific and common core lexis.

In addition to academic vocabulary, Paquot (2010:4–5) stresses the importance of English core vocabulary (high-frequency words) to academic prose (§3.4). For this reason, she included the first 2000 words in her Academic Keyword List (2010), together with 930 potential academic words. This marks a difference from Coxhead's (2000) Academic Word List (AWL), which contains only about 850 academic words. Paquot believes that 'a definition of academic vocabulary that excludes the top 2000 words of English is not very useful for productive purposes in higher education settings'. She argues for a 'function-based' definition of the term (Paquot 2010:9), supporting the findings of Hancioğlu et al. (2008).

In her study, Paquot (2010) used 10 ICLE sub-corpora of writing in English by learners with different European language backgrounds. These learners were all novice writers of English and of their own L1. She compared these texts with a sub-set of the BNC containing texts by specialists in the Humanities in order to determine how the learners' use of academic language differed from that of the expert writers. General features not specific to any particular language which emerged from Paquot's analysis (2010:125) were that the writing by learners in her corpus featured a limited lexical repertoire, with almost 50% of the words in her AKL being underused by the writers in the ICLE corpus. She also observed a lack of register awareness (cf. Read 2004; Hancioğlu et al. 2008), shown particularly by the overuse of lexical items that were more typical of speech than of expert writing; semantic misuse, for example of adverbials, or connectors, such as *on the contrary* and *on the other hand* that were overused in the ICLE and often used inappropriately; learners' writing sometimes contained too many connective devices, and these were often also used in an unmarked position, such as sentence initially (2010:150). In his study, Lake (2004:137) also found that international ESL EAP students had difficulty using prefabricated sequences (prefabs), often confusing the connective phrase *on the contrary* with *on the other hand*. These scholars' findings support those of Granger and Tyson (1996), who found an overuse of *for instance,* and also those of Henning (2006), who found that South African ESL students had trouble using connectors.

When she investigated exemplification in learner academic writing, Paquot (2010:125–126) found that exemplificatory lexical items, such as the adverbials *for example* and *for instance*, the noun *example* and the preposition *like*, were used significantly more often in the novice writing than in the professional corpus. She also observed that a limited set of clusters or prefabs such as *it depends* were learner-specific and 'massively overused' (2010:155). Granger (1998b) calls such expressions 'islands of reliability, citing Dechert (1984, cited in Granger 1998b:156), and these are also similar to Hasselgren's (1994) 'lexical teddy bears'. Paquot's findings also support Granger's (1998b:156) contention that learners' foreign-soundingness may be due in large part to their overuse, rather than their underuse, of prefabs (Paquot

2010:155). She found that 'learner writing is […] typically recognizable by a whole range of co-occurrences that differ from academic prose in quantitative and qualitative terms' (Paquot 2010:155).

These findings are particularly significant in the light of Read's (2007:120) observation that, at university level in particular, knowledge of academic register and vocabulary is crucial to success.

3.6 CONCLUSION

This chapter has provided a brief review of aspects of vocabulary research which are relevant to this study. Although focusing predominantly on breadth of vocabulary, the studies reviewed here have underlined the value of a study such as the one reported on in this thesis by pointing to the need for further research on the role of vocabulary knowledge, both breadth and depth, in academic performance. Studies of learners both at school level (Nation and Waring 1997; Scheepers 2003; Vermeer 2010) and, more importantly in the case of my study, at university level (Cooper 1999, 2000; Qian 2002; Morris and Cobb 2004; Paquot 2010; Milton and Treffers Daller 2013) thus contextualise my study. In the following chapter I review studies of MWUs and prefabricated language, as it is in terms of such phenomena that I explore a very specific aspect of undergraduate students' depth of vocabulary knowledge, as revealed in their writing.

Chapter 4

Formulaic language and multiword units

> The realisation that words act less as individual units and more as parts of lexical phrases in interconnected discourse is one of the most important new trends in vocabulary studies (Schmitt 2000:78).

4.1 INTRODUCTION

This chapter forms the final part of the literature review in this study. It discusses what is, in effect, the main focus of Phase 2, the delexical multiword unit (MWU). Accordingly, the chapter starts by providing some background to the growth in interest in the phenomenon of formulaic language in general in research over the last 30 years or so (§4.2). It documents the move from an awareness of the fixedness, or 'formulaicity' (Wray and Perkins 2000:1), of some language to the understanding expressed in the present study that formulaic language is not 'a single category' different from language freely generated by rules, but rather a term which covers all 'significant features of word combinations' (Howarth 1998a:25).

In the next section, the focus moves to collocation, a specific type of word combination which relates closely to the type of delexical MWU investigated in this study. Issues of definition of this construct and learning issues are discussed and influential studies reviewed (§4.3). In the last section, the focus narrows to the circumscription of the main construct under investigation in this study, the delexical MWU (§4.4). The definition of this combination as it is used here and the researchers who were most influential in this process are discussed, and the section ends with a review of the research studies on high-frequency verbs in general and on delexical uses of such verbs in particular that have informed this study.

By focusing on formulaic language in general and on MWUs containing delexicalised high-frequency verbs in particular, this chapter ties together the discussions in the previous two chapters: it reflects on the shift that has taken place in vocabulary studies, from discrete-item tests and the view of lexis as individual words towards the notion that words are integral parts of larger discourse, and it documents the influence that corpus studies have had on this move and the way in which corpus analysis has allowed researchers to isolate the type of delexicalised MWU studied in this chapter.

4.2 FORMULAIC LANGUAGE

As noted in Chapter 3 (§3.4), one of the most important findings to come out of studies of vocabulary in the last few decades, and from corpus research in particular, is that 'language is made up of not only individual words, but also a great deal of formulaic language' (Martinez and Schmitt 2012:299). This section provides some background to this development by discussing the growing awareness of the all-pervasive nature of formulaicity in language (§4.2.1), the challenges scholars have faced when trying to define formulaic language (§4.2.2) and the difficulties formulaic language poses for learners of English (§4.2.3).

4.2.1 The ubiquity of formulaic language

Research into formulaic language and MWUs has increased significantly in the last three decades, influenced in no small part by the increased use of computerised methods and corpora in linguistic studies (see Chapter 2). The focus of vocabulary studies has changed; a great deal of research has been devoted to explaining various lexical patterns (formulaic sequences, idioms, collocations, sentence stems, for example) based increasingly on corpus evidence. Software has been developed which allows researchers to lemmatise their corpora, to establish frequencies and generate concordances of specific words, and to identify collocational tendencies and many other aspects of their corpora. As Kaszubski (2000:2) observes, 'theory and corpus-based practice have shown that aspects of lexicon, phraseology and style are intertwined'.

Researchers have focused on MWUs and formulaic language, what Granger (1998b:145) calls prefabricated language or 'prefabs', and 'conventionalized language' (1998b:146), because of their frequency and because they are important to the native-like production of language (Cowie 1992; Fan 2009; Granger 1998b; Hunston and Francis 2000; Nesselhauf 2003; Wray 2002). Such chunks of language are also important to idiomaticity, which Kaszubski (2000:1) says is operationalised in the literature by the properties of 'non-compositionality of meaning and structure […] and conventionality and naturalness', or salience. In the words of Pawley and Syder (1983:91), 'fluent and idiomatic control of a language rests to a considerable extent on knowledge of a body of "sentence stems" which are "institutionalised" or "lexicalised"'. Cowie (1992:10) adds that 'it is impossible to perform at a level acceptable to native users, in writing or in speech, without controlling an appropriate range of multiword units'. And Renouf and Sinclair (1991:143) provide 'evidence in support of a growing awareness that the normal use of language is to select more than one word at a time, and to blend such selections with each other'.

Sinclair's work and the COBUILD project (Sinclair 1991) (see §2.2.2.2) made his approach to phraseology familiar to researchers in the field of English language and linguistic studies. Sinclair focused on recurrent co-occurrences of words in a body of texts and drew on Firth's concept of 'meaning by collocation' (Howarth 1998a:26). The more frequent such an occurrence, the more significant it was considered to be in

the language; for this reason, larger corpora that provided more data produced more reliable results. Howarth (1998a:26) sounds a word of caution here, however, in that 'a notion of significance based solely on frequency risks giving unwarranted emphasis to completely transparent collocations such as "have children"'. In other words, the researcher needs to establish criteria to identify exactly what sort of combination qualifies as a meaningful, non-compositional collocation; if it is to be useful, 'the notion of phraseological significance needs to take into account the differences between phraseological types and to consider how they are processed by native and non-native speakers and writers in production' (Howarth 1998a:27).

Sinclair's (1991) open-choice and idiom principle helped to concretise the growing awareness of the formulaic nature of language. This principle is based on the notion that, on the one hand, a language user has a huge choice of what words to use when saying or writing something, restricted only by the grammatical acceptability of the production, but that, on the other hand, there are also a large number of semi-preconstructed phrases, constituting single choices, which the language user could choose (Sinclair 1991:109–110). This is now supported by research findings: although studies differ hugely in the proportions of formulaic language they report, it is now generally accepted that there is far more lexical patterning and widespread collocation in language than was previously realised (Howarth 1998b; Hunston and Francis 2000). For instance, Altenberg (1998:102) estimates that over 80% of the words in the London-Lund Corpus 'form part of recurrent word-combination in some way or another'. While Moon (1998b), on the other hand, found that only 4% and 5% of the Oxford Hector pilot corpus of over 18 million words was made up of fixed expressions and idioms respectively; part of this discrepancy may lie in her more narrowly defined concepts. In a study conducted by Erman and Warren (2000) to explore the impact that prefabricated language has on the structure of a text and on the effort involved in encoding and decoding it, the authors found that there were large amounts of prefabricated language in both spoken and written texts (making up on average around half of the texts they investigated: 58.6% and 52.3% respectively). This 'makes it impossible to consider idioms and other multi-word combinations as marginal phenomena' (Erman and Warren 2000:29). These variations in the estimates of the proportion of formulaic language in any given corpus are a reminder of the complexities of formulaic language and its research: there is a host of definitions, many of which are superficially very similar.

Such formulaic expressions are often difficult for learners to understand, even when native speakers would regard them as fairly transparent (Martinez and Schmitt 2012). They also occur frequently in academic discourse, making them particularly important for learners of English in higher education contexts. Studies in phraseology such as those by Altenberg (1998), Gläser (1998) and Howarth (1998a, b) have revealed the blurring of the boundaries between grammar and lexis. Because of their functional importance, knowledge of MWUs is essential for pragmatic competence (Schmitt 2000:101). Shirato and Stapleton (2007), citing

McCarthy and Carter (2002, cited in Shirato and Stapleton 2007:409), claim that many high-frequency clusters occur with greater frequency than some common single words and pose great difficulties for ESL learners (see §4.2.3).

Thus, scholars are in agreement on the importance of formulaic language, but as they have used different criteria to establish exactly what makes something formulaic and may apply different terminology to these units, studies in this area are very difficult to compare (Wray 2002:28). For this reason, in the next section of this chapter an attempt is made to describe various researchers' conceptualisations of word combinations in general and to provide a clearer picture of the definitions and explanations these scholars have settled on.

4.2.2 Pinning down the phenomenon

As observed above, research in the last three decades or so has seen growing consensus on the formulaic nature of language, and the view that a great deal of text is made up of 'non-arbitrary and non-random phrases and patterns' (Kaszubski 2000:2) is generally accepted by scholars. With this consensus has come increased research and a plethora of terms and definitions for such patterns. Some studies have focused mainly on spoken data, and Wray (2000, 2002) is a particularly authoritative voice here. Studies in this area (Nattinger and De Carrico 1992; Schmitt and Carter 2004; Wray 2000, 2002; Wray and Perkins 2000) tend to focus on the pragmatic aspect of what are often termed formulaic sequences. Then there are those scholars who have focused more on written data, and in these studies a great deal of work has been done on lexical collocations. Such studies include those by Howarth (1998a, b), Granger (1998a, b), Altenberg and Granger (2001) and Nesselhauf (2003, 2004, 2005). Then there are many examples, for instance Sinclair's (1991) many studies and those by later scholars such as Biber et al. (1999) and Biber (2009), where both spoken and written data have been investigated.

Over the years, these studies of various manifestations of formulaic language have given rise to many different names and definitions for these combinations. In fact, Wray (2002:9) found more than 50 terms to describe these chunks of language. These include *prefabricated patterns* (Hakuta 1974); *chunks* (Peters 1983, cited in Shirato and Stapleton 2007:395); *lexical phrases* (Nattinger and De Carrico 1992); *recurrent sequences* (De Cock 1998; De Cock and Granger 2004); *prefabricated language* or 'prefabs' (Granger 1998b); *recurrent word-combinations* (Altenberg 1998); *lexical bundles* (Biber et al. 1999); *multiword units* (MWUs) (Schmitt 2000); *formulaic sequences* (Schmitt and Carter 2004; Wray 2000, 2002); as well as *idioms*, *collocations*, *formulas*, *formulaic speech*, *prefabricated routines*, and *ready-made utterances*.

However, as Wray (2000:464) points out, these are not simply different names for the same phenomenon: 'A full appreciation of what formulaic language is requires us to recognise that we are not dealing with a

single phenomenon but rather with a set of more and less closely related ones, across different data types', including data from first language learners, second language learners, adult native speakers and even those with linguistic disabilities, and including both spoken and written data. Thus, in the light of developments in the last few decades, Weinert's (1995:182) assertion that 'while labels vary, it seems that researchers have very much the same phenomenon in mind' is disputed by both Howarth (1998a:25) and Wray (2000). Howarth feels that Weinert ignores the problems associated with the categorisation of these combinations of words, and applies these terms 'too loosely' to a range of chunks of language which are possibly significantly different from each other (Howarth 1998a:25), while Wray's argument echoes Howarth's words (see §4.1) that formulaic language is not 'a single category', but one which covers all 'significant features of word combinations' (Howarth 1998a:25).

Howarth (1998a) believes that several phraseological features or processes can be identified in the terms which Weinert (1995:182) uses interchangeably when she refers to formulaic language as 'multi-word […] or multi-form strings […] which are produced or recalled as a whole chunk, much like an individual lexical item' (Weinert 1995:182). For instance:

- Weinert's 'multi-word or multi-form strings' could be interpreted in terms of 'the formulaic nature of expressions (that is, conventional "form-meaning pairings" (Pawley and Syder 1983) that become institutionalised in language' (Howarth 1998a:25);
- Weinert's description, 'recalled as a whole', suggests memorisation (which Howarth [1998a] regards as a characteristic of the individual language user);
- Weinert's 'as a whole chunk' suggests lexicalisation – 'when a multiword item becomes stored and processed unanalysed as if it were a simple lexical item' (Howarth 1998a:25);
- Howarth adds a fourth feature which is not mentioned by Weinert but which he believes is implicit in her description, that of fixedness ('fixed expressions', e.g. Moon 1992) (Howarth 1998a:25).

There is considerable overlap between the bulleted categories in terms of most features applying to one and the same piece of formulaic language. This suggests that Howarth (1998a) and Wray (2000) are clearly more concerned to isolate and categorise properties than Weinert (1995); thus, although she was aware of the features that are taken up by Howarth (1998a), she tends to use the terms interchangeably at times. So Howarth (1998a) is reluctant to say, as Weinert (1995) does, that formulaic language is one all-encompassing category. He also believes that properties of formulaic language are 'gradable' (Howarth (1998a:25–26) – some combinations that are used by children as unanalysed chunks will become analysable in adult speech, for instance, while some 'unanalysed' chunks may be produced unanalysed by a native speaker (NS), but compositionally by a non-native speaker (NNS). It is clear, then, that part of the difficulty in providing a comprehensive definition of these combinations lies in their very diversity (Schmitt and Carter 2004:2). They differ in length and function, they may be totally fixed in form or may include

variable 'slots' into which certain words can be inserted. Criteria for defining these MWUs include institutionalisation, fixedness and non-compositionality (Howarth 1998a; Moon 1997:44; Schmitt and Carter 2004:2). Howarth (1998a:27) believes that 'the notion of phraseological significance needs to take into account differences between phraseological types and to consider how they are produced by native and non-native speakers and writers in production' – something which this study sets out to do with regard to delexical MWUs.

Drawing on Russian lexicography in particular, Howarth (1998a) established a set of features as a basis for the categorisation of word combinations. For Howarth (1998a) it is vital that researchers use the criteria he lists (as indicated below in Figure 4.1) to distinguish between categories of combination, avoiding the use of the very broad term 'formulas':

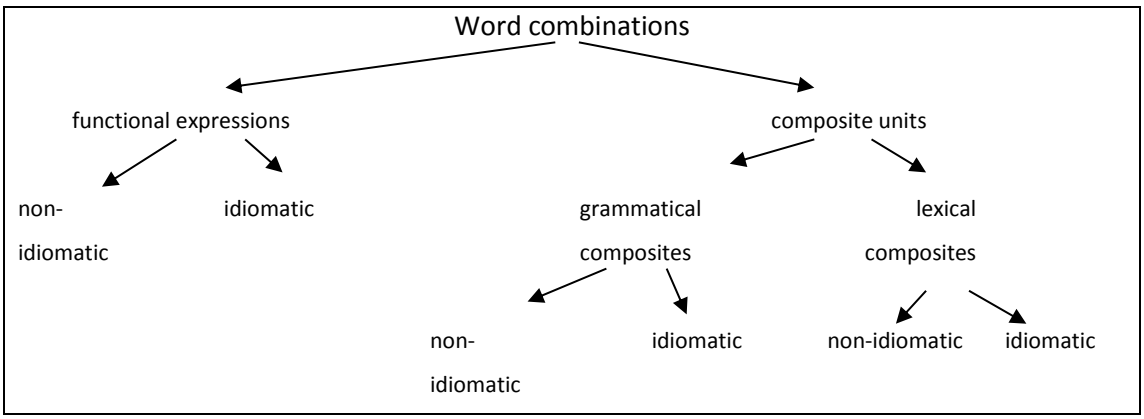


Figure 4.1: Phraseological categories (Howarth 1998a:27)

Howarth's further categorisation of composite units into lexical and grammatical composites revealed to him that this was not 'a simple two-way division' but 'a continuum derived from the application of such criteria as restricted collocability, semantic specialization, and idiomaticity, each of which is gradable' (Howarth 1998a:28).

Later, Wray chose as her definition for these combinations a term which she believes does not have previous 'baggage' (2002:9) – *formulaic sequences*. *Formulaic* has 'associations of "unity" and of "custom" and "habit", while *sequence* indicates that there is more than one discernible internal unit, of whatever kind'. She defines her term thus:

> A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray 2002:9).

Like Wray (2000, 2002), Schmitt and Carter (2004:4) also settle on the term *formulaic sequences* because this describes the two aspects of the phenomenon they wish to emphasise: (a) these sequences are not just any sequence of words but phrases; and (b) they are lexical items just as words are, with the same properties that words would have. They list characteristics of these formulaic sequences rather than identifying specific criteria which define them, as they are so diverse. These include:

- *'Formulaic sequences appear to be stored in the mind as holistic units, but they may not be acquired in an all-or-nothing manner'* (Schmitt and Carter 2004:4).

  These sequences fall on a continuum, with idioms at the most opaque end and transparent sequences at the other. This may affect the ease with which they are learnt. This is similar to what Howarth (1998a, b) found when he categorised verb and noun collocations into three main levels of restrictedness (free collocations, restricted collocations and idioms). These categories are illustrated in Howarth's scheme, as reproduced below:

	Free combinations	**Restricted collocations**	**Figurative idioms**	**Pure idioms**
Lexical composites Verb + noun	*Blow a trumpet*	*Blow a fuse*	*Blow your own trumpet*	*Blow the gaff*
Grammatical composites Preposition + noun	*Under the table*	*Under attack*	*Under the microscope*	*Under the weather*

Table 4.1: Howarth's (1998a:28) collocational continuum

  Schmitt and Carter (2004:6–7) also observe that these sequences may allow some flexibility in their composition.

- *'Formulaic sequences can have semantic prosody'*

  Those learners who are proficient in the language will understand this – they will know that certain words occur typically in a particular structure, such as 'SOMETHING UNDESIRABLE is/are rife in LOCATION/TIME' (Schmitt and Carter 2004:8). This is the notion of idiomaticity – a sense of the 'salience', as Granger (1998b) puts it, and an awareness of the 'conventionality and naturalness of some expressions' (Kaszubski 2000:1).

- *'Formulaic sequences are often tied to particular conditions of use'* (Schmitt and Carter 2004:9).

  These include functions such as speech acts – apologising, requesting, etc. Formulaic sequences are particularly useful in maintaining social interaction, that is, 'small talk' (Schmitt and Carter 2004:5–10). This criterion also distinguishes their definition from Howarth's notion of collocation, which does not involve communicative function. In fact, Howarth suggested two fundamental phraseological categories, functional expressions, which play a role in discourse, such as gambits and catchphrases

(similar to Schmitt and Carter's [2004] 'speech acts'), and composite units, including collocations, which have a syntactic function (described in the table above). Both these types can be idiomatic or non-idiomatic (Howarth 1998a:27).

Biber argues that there are 'two underlying linguistic constructs: multi-word lexical collocations versus multi-word formulaic sequences' (2009:277), and 'two major parameters' (2009:301) necessary for the description of formulaic language in English: 'collocations versus multi-word formulaic sequences, and fixed continuous sequences versus formulaic frames (with internal variable slots)'. He describes how he took 'a radical corpus-driven approach to investigate the ways in which high-frequency sequences of words pattern in terms of fixed and variable slots' (2009:277), differing from studies such as Renouf and Sinclair (1991), which included a corpus-driven element in that the researchers 'pre-selected' (Biber 2009:276) the collocational frameworks they searched for. Instead, Biber identified the patterns that occurred most commonly in his corpus, determining the different ways in which those patterns were variable or fixed, and contrasting the overall patterns used in speech and writing (Biber 2009:291).

Biber (2009) analysed two corpora, a corpus of American English conversation comprising 4.5 million words, and one of academic prose, containing 5.3 million words. Considering only four-word sequences, he found that some of these sequences were fairly formulaic (such as *on the other hand*) while others were more variable, such as *are likely to be* (Biber 2009:292). His analysis revealed unexpected trends: the preference in conversation for continuous fixed sequences as opposed to the preference in writing for formulaic frames with internal variable slots (Biber 2009:300). These findings supported Biber's earlier research into 'lexical bundles' (Biber 2006), leading him to the conclusion that formulaic language is extremely important in both conversational and written academic discourse, 'but it is realised in very different ways linguistically' (Biber 2009:301), making the grammar of speech fundamentally different from the grammar of writing.

Biber (2009:275) focused on 'multi-word formulaic sequences (incorporating both function and content words)' rather than 'multi-word lexical collocations (combinations of content words)' (although his definition of a collocation is somewhat narrower than the one I use), supporting the implication that there are two strands to the study of formulaic word combinations, and that the use of such combinations differs in speech and writing. In the light of this, I focus in §4.3 on collocations. This is the area into which my own study best fits, as I investigate the use of MWUs in academic writing, and most work in this area has been done with written data. However, references are made to studies dealing with spoken language and multi-word formulaic sequences where necessary. For this reason, too, the next major section (§4.3) is devoted to defining the notion of collocation and to research relating to these combinations. Before this, however,

difficulties learners experience when they encounter MWUs and formulaic language in general are discussed.

4.2.3 Challenges posed by formulaic language to learners of English

Whatever we choose to call these word combinations, it is clear from the studies mentioned above and elsewhere in this thesis that, far from consisting of an infinite number of creative utterances, a native speaker's spoken and written output is made up of a significant number of prefabricated, multiword items. Wray (2002:13), for instance, believes that formulaicity is 'all-pervasive in language data' and that words belong with other words at the most basic level, what Sinclair (1991:110) refers to as 'unrandomness'. Many researchers (e.g. Erman and Warren 2000; Sinclair 1991; Wray 2002) believe that neither an analytical nor a holistic process alone can 'accommodate both the linguistic competence of the ideal native speaker and listener and the idiomatic choice of one grammatical string over another' (Wray 2002:15). This may in part account for what Pawley and Syder (1983:193) call the 'puzzle of nativelike selection' – a native speaker's utterances are both 'grammatical' and 'nativelike', but only a 'small proportion' of grammatically well-formed sentences are native-like, that is, 'readily acceptable to native informants as ordinary, natural forms of expression', and it is these which native speakers produce in their speech and writing. Harwood (2002:140) agrees that speakers – and of course learners of the language – use both what he calls 'automatized', prefabricated expressions, as well as creative, language: what Sinclair (1991) refers to as 'idiom' and 'open choice' components respectively (see §4.2.1). Such word combinations are 'inextricably related to style – the appropriateness and/or naturalness of selection and co-occurrence of items' (Kaszubski 2000:2). This is a vital aspect of academic success; university students in particular must be able to master the appropriate academic style and register.

Formulaicity in language, particularly in speech, is ubiquitous because it makes processing much easier and quicker. The all-pervasive nature of formulaicity that corpus studies have revealed reflects the processing principle of 'economy of effort' (Sinclair 1991:110; Wray and Perkins 2000:11). '[I]t is our ability to use lexical phrases' that allows us to be fluent (Nattinger and De Carrico 1992:32); it is to our advantage to be fluent because we can convey more information, speed up processing of input, gain the centre of attention in a conversation more easily, and cope better with our short-term memory limitations. Wray (2002:18) sums it up as follows: the analytic system allows us to create new expressions and to interpret new input we receive. The holistic system reduces the processing effort – but this is much more than simply a system to retrieve idioms.

Among second language learners, formulaic sequences are relied on initially as a quick way to communicate, particularly in speech (Schmitt and Carter 2004:11; Wray 2002:ix). This may aid integration in a peer group and lead in turn to increased language output and further language learning: '[t]here is little

doubt that the automatic use of formulaic sequences allows chunking, freeing up memory and processing resources' (Kuiper 1996, cited in Schmitt and Carter 2004:12; Ellis 1996, cited in Schmitt and Carter 2004:12). These resources can then be used to cope with conceptualising and meaning. Wray (2002:ix) supports this, noting that native speakers tend to use formulaic language as 'an easy option in their processing and/or communication'. In the early stages of both first and second language acquisition learners rely a great deal on formulaic language but, paradoxically, formulaic language seems to be the biggest obstacle to achieving native-like fluency for intermediate and advanced L2 learners 'because the learner lacks the necessary sensitivity and experience that will lead him or her unerringly away from all the grammatical ways of expressing a particular idea except the most idiomatic' (Wray 2000:463). The 'most idiomatic way', very often, involves an MWU.

Many authors report on the challenges MWUs may present to successful language learning, particularly at advanced levels where students are expected to adhere to the constraints of academic genres. Nesselhauf (2003, 2004, 2005), Granger (1998b), Howarth (1998b), Fan (2009), Farghal and Obiedat (1995), Shirato and Stapleton (2007), Stubbs (2001), Altenberg and Granger (2001) and Laufer and Waldman (2011) have all shown the difficulties learners experience with collocations. For instance, Shirato and Stapleton (2007:409) found in their study that Japanese EFL learners' conversation revealed only a limited use of high-frequency words and a lack of multiword clusters, making their speech sound blunt and pedantic (2007:410). They suggest that multi-word clusters are neglected in Japanese teaching to the detriment of the learners. Like Shin and Nation (2008), they stress the importance of high-frequency collocations.

These difficulties are compounded by the all-pervasive nature and the variety of these combinations. For instance, Hyland (2008) concluded that certain 'bundles' were specific to or predominated in particular disciplines and discourses and their use and naturalness in these contexts signposted 'competent participation' in these writing and speech communities (2008:5). As Hyland observes, writers' failure to use these clusters of words might result in a lack of fluency and native-like idiom. He puts it thus:

> … gaining control of a new language or register requires a sensitivity to expert users' preferences for certain sequences of words over others that might seem equally possible (Hyland 2008:5).

The sections above have provided some background to research on types of formulaic language and MWUs. The importance of these units to learners of the language and the challenges this presents was also touched on. This latter aspect is covered in greater detail below, where relevant research studies are discussed (§4.3.2). In the following sections the discussion focuses on collocation where definitions of this type of word combination are discussed, and ESL issues related to collocations are considered.

4.3 COLLOCATION

In this section collocation in general is considered, along with how various scholars have explained and defined this type of word combination. The reason for this focus on collocations is that the delexical MWUs investigated in this study fit best within the term *collocation*. Schmitt (2000:76) believes that one of the major insights to have been gained from the study of corpora concerns collocations, which have been investigated almost exclusively through corpus evidence (Schmitt 2000:68). Simply put, collocation is the tendency of two or more words to co-occur in discourse. Sinclair's (1991) distinction between the open-choice principle and the idiom principle is relevant here (see §4.2.1 above). The open-choice principle is the notion that language is creative, and in most instances a wide variety of words could be inserted into a given 'slot'. However, complementary to this freedom of choice, Sinclair notes that language also has a systematicity that restricts vocabulary choice in discourse (Schmitt, 2000:7). Much of this systematicity is strictly linguistic – there are regularities in how words co-occur with each other; collocation is one term that covers this notion.

4.3.1 Defining collocation

As Bahns (1993) points out, collocation is expressed differently by various scholars: he makes a distinction, between grammatical and lexical collocations, similar to Kjellmer's (1991) below and to Howarth's (1998a, b) (see §4.2.2). As Bahns (1993) defines them, grammatical collocations comprise noun/adjective/verb + preposition or a grammatical structure such as an infinitive or a clause (*by accident*, *to be afraid that*), while lexical collocations are combinations of noun/adjective/verb/adverb such as *inflict a wound*, *withdraw an offer* (Bahns 1993:57).

In a fairly early study, Kjellmer (1991:112) notes that 'a large part of our mental lexicon consists of combinations of words that customarily co-occur. The occurrence of one of the words in such a combination can be said to predict the occurrence of the other(s)'. Like many other researchers, however, he believes that categorisation is difficult and that boundaries between any categories we devise will necessarily overlap at times. He provides a detailed discussion of 'fossilised phrases' ('sequences where the occurrence of one word almost unequivocally predicts the occurrence of another' [Kjellmer 1991:112–113], such as *arms akimbo*, *from afar*) and 'semi-fossilised phrases' (where one word predicts a limited number of words, such as *by and by/by and large*, *moot point*/*moot question*). He believes that idioms belong chiefly to this latter category and suggests examples such as *have a weak/soft spot for*, *get off on the right/wrong foot* (1991:113), noting however that in his definition a sequence representing a grammatical pattern is 'admitted as a collocation only if it meets certain lexical conditions' because collocations are lexically selected (1991:118). Kjellmer uses the term collocation to refer to 'structured patterns which recur in identical form' (1991:116). In his study, he examined and compared a sample of prose to 'a corpus of collocations' (1991:116) extracted from the Brown Corpus and made up of 85 000 collocational types. He

found that 'even the words occurring between those [fixed] combinations constitute groups whose form and order are likely to be conditioned in varying degrees by patterns of collocability' (1991:120). But these collocations are not always a clearly distinguishable category – Kjellmer, like Howarth (1998a, b) and Schmitt and Carter (2004), talks of a 'collocational continuum' (1991:121) of established collocations, such as *subject to* and *all kinds of*, at one end, and at the other end 'sequences of doubtful cohesion' (1991:121) such as *possible criteria* and *linguistic situation*, where free choice, or open choice (Sinclair 1991), and a more productive element operate.

Renouf and Sinclair (1991) also explored collocations, but of a type that differs from Kjellmer's (1991) definition above. Their focus was on 'frameworks' made up of a 'discontinuous sequence of two words, positioned at a one word remove from each other' (Renouf and Sinclair 1991:128). These frameworks rely for their grammatical 'self-standing' (1991:128) on the word which comes between them. Examples of the frameworks they investigated are a *+ ? + of (a kind of)*, *an + ? + of (an act of)*, *be + ? + to (be able to)*, and *too + ? + to (too late to)*.

Renouf and Sinclair (1991) based their investigation of these frameworks on a one-million word corpus of spoken British English and a 10-million word corpus of written British English, both sections of the Birmingham Corpus of English Text, part of the COBUILD corpus. They believe that their findings showed that 'two very common grammatical words, one either side, offer a firm basis for studying collocation' (1991:143). Their findings revealed that these frameworks were statistically important in their corpora. Evidence of this statistical significance lies, they believe, in the 'high type-token ratio' (1991:143); that is, their study revealed, on average, 'a very high rate of recurrence of types in proportion to the number of framework tokens' (1991:130), indicating that these frameworks are 'highly selective in their collocates'(1991:130) and that 'the choice of word class and collocate is specific, and governed by both elements of the framework' (1991:143). In this way, Renouf and Sinclair presaged the growing body of research which supports the notion that it is normal for language users to 'select more than one word at a time, and to blend such selections with each other' (Renouf and Sinclair 1991:143).

As Schmitt and Carter (2004) note, multiword formulaic sequences are not necessarily continuous fixed sequences of words. Rather, as Renouf and Sinclair (1991:129) showed in their study of collocational frameworks, they may be made up of different pairings of high-frequency grammatical words, such as *a + ? + of*, *be + ? + to* (as in *a lot of*, *be allowed to*), and the word that comes between them, which they refer to as the collocate; the two grammatical words in these examples are also in a relationship with one another, but this would more usually be regarded as colligation. There may also be discontinuous frameworks comprising fixed slots filled by a single function word together with variable slots filled by many different content words, such *as a/n + ? + of + the* (Biber 2009:290–291).

In his study, Howarth (1998a, b) also distinguishes lexical collocations from grammatical collocations. Lexical collocations consist of two open-class words such as V + N or Adj + N, while grammatical collocations are combinations of N/V/Adj + closed class word such as a preposition. Schmitt (2000:77), much like Howarth (1998a, b), also suggests different categories of collocation and identifies two basic kinds: (1) grammatical/syntactic collocations and (2) semantic/lexical collocations. In the first type, a dominant word 'fits together' with a grammatical word (like Howarth's [1998a] grammatical composites). These are typically a verb, noun or adjective followed by a preposition – *abide by*, *access to*. In the second type (like Howarth's lexical collocations), two basically 'equal' words such as noun + verb (*the ball bounces*), verb + noun (*spend money*), adjective + noun (*cheerful expression*) combine, with both words in the combination contributing to meaning. However, Schmitt notes (2000:78), like Renouf and Sinclair (1991), that it has become increasingly clear that there are also collocations made up of strings of words. Nattinger and DeCarrico (1992), for instance, believe that one needs to look more than five words away from the core word to find every collocational relationship.

Schmitt (2000:89) observes that although it is not clear how collocational knowledge is acquired, it seems to be relatively difficult to achieve, and collocational ability is often one of the aspects of speech or writing that distinguishes native speakers from non-native speakers. This has been borne out in other studies such as those by Bahns (1993), Granger (1998b) and Nesselhauf (2003, 2004, 2005). Howarth, too, believes that collocations

> are not optional stylistic adornments on the surface of text; they are essential for effective communication, and their use by non-native writers is a clear sign that these learners have made an essential adjustment to the academic culture they are entering (Howarth 1998b:186).

This is further motivation for the focus on these combinations in my study.

In an investigation of collocational use by NNSs and NSs, Howarth (1998a) took his NS data from sections of the LOB, while his NNS data comprised 10 essay assignments by postgraduates completing an MA in applied linguistics. He subsequently compared the use of lexical collocations comprising transitive verb + object noun combinations in formal written English. He chose these specific combinations because he viewed them as important as far as the propositional content of an academic argument is concerned. 'Much of the distinctive procedural vocabulary of academic disciplines can be found in such predicate structures as *make a claim*, *reach a conclusion*, *adopt an approach*, *set out criteria*' (Howarth 1998a:34), many of which are delexical MWUs like the ones analysed in the present study. Howarth believes that the mastery of combinations such as these can be a sign of NS competence and 'a useful indicator of degrees of proficiency across the boundary between native and non-native competence' (Howarth 1998a:38), something which my own study hopes to illustrate.

Howarth (1998b:169–170) categorised collocations into five levels, from least to most restricted. The levels which pertain to the MWUs discussed in my study are:

- Level 2, where some substitution of both verb and noun is allowed, that is, a small group of nouns can be used with the verb as it is used in the particular sense and there is a small number of synonymous verbs, e.g. *introduce*/*table*/*bring forward a bill*/*an amendment* (Howarth 1998b:169);
- Level 3, where some substitution of the verb is allowed but the choice of noun is completely restricted, that is, only the specific noun can be used with the verb in the given sense, and there is a small group of synonymous verbs, e.g. *pay*/*take heed* (Howarth 1998b:170); and
- Level 4, where the choice of the verb is completely restricted but some substitution in choice of noun is allowed, that is, a small group of nouns can be used with the verb in the particular sense but there are no synonymous verbs, e.g. *give the appearance*/*impression* (Howarth 1998b:170).

Level 1 covers free combinations while at Level 5 the choice of both verb and noun is completely restricted; such combinations are conventionally termed idioms.

As is the case in most research in this area, Nesselhauf (2003, 2004, 2005) also grappled with issues of definition. She explains that 'collocation' is used in her study 'in a phraseological rather than in a frequency-based sense' (2003:224), in other words, to indicate a particular combination of words rather than words which co-occur in a particular span (cf. Sinclair 1991). Nesselhauf uses 'only the most widely accepted defining criterion for collocations […] namely arbitrary restriction on substitutability' (2003:224–225). She thus makes her distinction between free combinations (Cowie 1994, cited in Nesselhauf 2003:225; Howarth 1998a), where possible substitution relies on 'semantic properties', and collocations, in which restriction on substitution is somewhat 'arbitrary' (2003:225). Thus her notion of 'restricted sense' (Nesselhauf 2003:225) is at the core of her definition of collocations and, like Howarth (1998a, b) before her, she has developed a system of restriction applied to verb-object noun combinations. The sense of the verb (or noun) is regarded as restricted if one of the following applies: (1) the sense of the verb (or noun) is so specific as to allow combination with only a small set of nouns (or verbs), or (2) the verb (or noun) cannot be used in the particular sense with all nouns (or verbs) that are syntactically and semantically possible (Nesselhauf 2003:225). She provides as an example the combination *read a newspaper*, where substitutions such as *drink a newspaper* or *read water* are unacceptable as the verb 'read' requires a noun with a 'semantic property' of 'containing written language' while *drink* requires a noun with a 'semantic property of liquid'. In a combination such as *reach a decision*, however, *decision* can be replaced by several other nouns with the meaning 'a particular aim', such as *conclusion* or *goal*, but not *aim*, 'a somewhat arbitrary convention of the language' (Nesselhauf 2003:225).

On the basis of her definition of restriction, Nesselhauf defines three classes of word combination:

- Free combinations – in which both verb and noun can be freely combined as their senses are unrestricted (*want a car*);
- Collocations – the noun's sense is unrestricted, but not the verb's. In the particular sense in which it is used, the verb can be combined with certain nouns only (*take a picture*/*photograph* but not *take a movie*/*film*);
- Idioms – the sense of both noun and verb is restricted. Substitution is impossible or extremely limited (*sweeten the pill*) (Nesselhauf 2003:226).

Nesselhauf's (2003) definition has informed my own definition of the combinations I have focused on in this study.

In the following section, studies on issues related to collocation in the context of ESL are explored.

4.3.2 ESL issues related to collocation

Scholars mentioned above, such as Howarth (1998a, b) and Nesselhauf (2003, 2004, 2005), have pointed to the difficulties learners have in mastering these restricted or semi-restricted combinations or collocations, such as those considered in this study. Teachers also find them difficult to explain because they are not all learnt as 'inflexible wholes' (Howarth 1998a:38), as idioms or more restricted collocations would be. But although they are not always 'fully lexicalised, they are quite institutionalized, and therefore form a part of the stock of complexes that help to mark a piece of writing as natural and proficient' (Howarth 1998a:38). As Nesselhauf (2003) also demonstrates, although some collocations can be 'predicted by analogy' they are 'arbitrarily blocked by usage' (Howarth 1998a:37). Howarth (1998a, b) and Nesselhauf (2003) believe that difficulty in selecting appropriate co-occurring items is part of the much more general phenomenon of arbitrary lexical and/or grammatical restriction. Collocates, or the words with which a particular word keeps company, are often *not* arbitrary at all but are in fact what Howarth (1998a, b) refers to as part of 'restricted' collocations (such as *pay/take heed* and *give the appearance/impression*). In his study, Howarth (1998a, b) observed that the largest group of errors made by NNS writers in his data fell into the category of 'restricted collocations'. He found that the proportion in the NS corpus of the two categories of restricted collocations and idioms which could be regarded as conventional was 38%. This finding is an indication of the collocational density and the degree of stylistic conventionality (an important aspect at university level) of academic writing and also allows comparison with analyses of other registers. The NNS writers, on the other hand, produced a much lower density of conventional combinations (25%), which suggests either a generally lower level of knowledge of collocations, or a lack of awareness of how to use them appropriately, or both, a finding confirmed by Granger (1998b) (Howarth 1998a:34–36).

It seems that learners do not always understand the notion of restriction – the fact that, although a sentence might be grammatically correct, a native speaker would never use it – but 'knowledge of arbitrary restriction on collocation is required' to conform to the expectations of the academic community and to comply with the conventions of academic style (Howarth 1998b:162; Nesselhauf 2003, 2005). Learners need to understand that restricted collocations make up a significant part of a typical native speaker's production in both speech and writing. Howarth makes reference in this regard to Granger's finding that learners have an 'underdeveloped sense of collocational "salience"' (Granger 1998b:152). This is an aspect which I investigate in my own study, in order to go some way towards explaining students' difficulties in conforming to stylistic conventionality in their writing. This is an area in which teachers themselves require guidance and I hope the study will shed further light on this aspect of language learning.

In another study focusing on issues of formulaic language in ESL, Sylviane Granger's work at Louvain on word combinations was part of the International Corpus of Learner English (ICLE) project, the aim of which was to gather and computerise a corpus of EFL writing from learners of various mother-tongue backgrounds (Granger 1998b:146). Employing what she terms 'Contrastive Interlanguage Analysis' (CIA), Granger compared a corpus of native English writing and a similar corpus of writing by advanced-level French-speaking learners of English taken from the ICLE. Her hypothesis was that learners would make less use of prefabs, or conventionalised language, in their writing than their NS counterparts, given that the use of such language is 'universally presented as typically native-like' (Granger 1998b:146). She hypothesised that learners would make greater use of the 'open choice' principle (Sinclair 1991).

In this investigation, Granger (1998b:154) makes a distinction between two types of word combination, formulae or 'sentence builders', which have a more pragmatic function, and collocations, which have a syntactic function. This again supports the notion that there are two fundamental types of MWU or prefabricated expression, namely, those with a pragmatic function (such as *good morning*, *as I was saying*) and which are used particularly in spoken language as conversational gambits, for instance (Wray's [2000, 2002] formulaic sequences, for example), and those with a lexical function (such as Granger's [1998b:152] examples *blissfully happy* and *highly significant*), what Howarth (1996, 1998a, b) would call collocations. Granger (1998b:146–7) explains the latter as examples of a 'linguistic phenomenon' where a particular word is found together with an item other than its synonyms because of restriction of usage rather than of syntax or conceptual meaning.

I will focus here on Granger's investigation of collocations in her corpora as this is most pertinent to the type of MWUs I investigate in this study. She uses Van Roey's (1990, cited in Granger 1998b:146–147) definition of collocation: 'the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its "synonyms" because of constraints that are not on the level of syntax or

conceptual meaning but on that of usage', or what Howarth (1998a, b) and Nesselhauf (2003, 2004, 2005) would argue are 'arbitrary'.

Granger (1998b) selected one category of intensifying adverbs: amplifiers which end in *-ly* and function as modifiers (*perfectly* normal, *closely* linked, *deeply* in love), because these involve semantic, lexical and stylistic restrictions and range from restricted collocability – *bitterly cold* – to more open collocability – as in *completely different/new/free* (Granger 1998b:147). Her analysis of the two corpora revealed a 'statistically very significant underuse of amplifiers in the NNS corpus, both in number of types and of tokens' (1998b:147). But learners overused *completely* and *totally* in a way which suggested that they may have regarded these words as 'safe-bets' (Granger 1998b:148) (what Hasselgren [1994] would call 'lexical teddy bears') in all sorts of combinations, while *highly* was underused. The fact that the words *completely* and *totally* have direct translations in French, or were 'lexically congruent' (Granger 1998b:150; cf. Bahns 1993:59) may also have accounted for this overuse (Granger 1998b:148), and this predicts Nesselhauf's (2003) findings of the influence of congruity between L1 and L2.

When Granger categorised amplifiers either as 'maximisers' (expressing the highest degree, such as *absolutely, entirely, totally*) or as 'boosters' (expressing a high degree only, such as *deeply*, *strongly*, *highly*), she found an underuse of boosters by NNSs significant enough to explain the general underuse of amplifiers remarked on above (1998b:149). Boosters made up two-thirds of the amplifiers in the NS corpus, compared to only a third in the NNS corpus. Granger found that boosters which were used exclusively by NSs fell into two categories: stereotyped combinations (*acutely aware*, *keenly felt*, *painfully clear*) and creative combinations such as *ludicrously ineffective* and *ruthlessly callous* (Granger 1998b:150). Both types were significantly underused by NNSs; thus she found that although NNSs were using collocations, they were underusing native-like collocations and used 'atypical word-combinations' (1998b:152). This may be due to an undeveloped sense of salience for what constitutes a significant collocation (Granger 1998b:152; Howarth 1998a). This interpretation is supported by the finding of her investigation of sentence-builders in the study that 'while the foreign-soundingness of learners' production has generally been related to the *lack* of prefabs, it can also be due to an excessive use of them' (Granger 1998b:155).

Nesselhauf (2003), like Granger (1998b), found that congruity between L1 and L2 exerted an influence on the use of collocations; she investigated the use of verb-object noun collocations (see §4.3.1), such as *take a picture*, *draw up a list*, by advanced German-speaking learners of English, and the difficulties they experienced. She supports the contention that collocations are important to second language learners as they enhance not only accuracy (particularly important for students who wish to succeed academically in the second language) but also fluency (Boers, Eykmans, Kappel, Stengers and Demecheleer 2006; Gledhill 2000, cited in Laufer and Waldman 2011:648; Nattinger and De Carrico 1992 Stubbs 2001; Wray 2002). She

took her data, 32 essays written by German-speaking students of English in their third or fourth year of university, from the German sub-corpus of ICLE. Her method was to manually extract all examples of verb-noun combinations from the data, to classify these according to the degree of restriction and to evaluate the acceptability of these combinations in English (Nesselhauf 2003:227).

Nesselhauf's (2003) findings revealed that even advanced learners of English struggle in the area of collocation, most errors being the wrong choice of verb, which she says is not surprising as in her definition of collocations the sense of the verb is restricted. But while the degree of restrictedness certainly contributes to the learner's ability to produce combinations which are acceptable to the native speaker, the first language exerts a much greater influence on the acceptability of the word combinations a learner produces – if the combination is directly translatable from English into, in this case German, the learner is much more likely to produce an acceptable combination of words (Nesselhauf 2003). She thus suggests that certain criteria be considered when deciding which collocations should be taught:

- They should be acceptable to native speakers.
- They should be frequent in any material which a student might be expected to encounter.
- Non-congruent combinations and those combinations which are less restrictive should receive particular attention.

Another interesting detail which the study revealed was that, contrary to common belief, restrictedness was less important than Nesselhauf had expected, but verbs in less restricted combinations presented learners with considerable challenges, particularly those which were not congruent with their first language (Nesselhauf 2003), such as in the case of the incorrect *make homework* (correct *do homework*), which was influenced by the German expression *Hausaufgaben machen* (Nesselhauf 2003:235).

Nesselhauf's (2003) findings support what Bahns (1993:56) found, namely, that learners encounter difficulties in learning the collocational properties of new vocabulary, especially where there is no direct translational equivalent in their mother tongue. It is these collocational errors in particular which mark a learner's speech or writing as foreign. For instance, in a supporting study by Bahns and Eldaw (1993), the authors found that German advanced EFL students' knowledge of collocations had not kept pace with their knowledge of vocabulary in general. Bahns (1993) believes that collocational errors, which impede the learning of truly idiomatic English, arise most often from what he calls the 'hypothesis of transferability' (Bahns 1993:61; see also Nesselhauf 2003, 2004, 2005), and can be traced back to the influence of students' L1. It is also probable that the exposure factor has a role to play here: students who lack exposure to texts in the target language do not develop their vocabulary through reading. This is an aspect of

particular importance to my own study, where many students come from print-poor environments and may lack a reading habit (see §1.2).

As in Nesselhauf's (2003) study, the corpora used in my study are relatively small when compared to corpora such as the BNC or LOCNESS. But many of the studies mentioned in this thesis (Fan 2009; Flowerdew 2001; Nesselhauf 2005) have been done recently using smaller, teacher/researcher collected learner corpora, with interesting results. These are particularly useful in identifying learners' under- or overuse of both spoken and written vocabulary, and can also reveal how far NNSs deviate from NS norms.

One such study was conducted by Fan (2009) in an attempt to form a clearer understanding of the competence in L2 collocational use of ESL secondary school leavers in Hong Kong. Fan (2009:112) defines collocations as 'the co-occurrence of two or more words within a short space of each other in a sentence context, involving lexical or grammatical words'. Her study examined collocations of 'different degrees of fixity and transparency in syntax and meaning' (Fan 2009:112). Fan believed that by using what she calls a 'loose definition' she would be able to form a deeper understanding of the general use of collocations by learners and some of the problems they encounter.

Fan looked at the learners' actual performance in a writing task, using a NS corpus of 60 'expert' essays from a specially arranged essay competition at a school in England and a NNS corpus compiled from essays written under examination conditions by 60 Cantonese-speaking Hong Kong students (2009:113). Fan identified all lexical words (i.e. nouns, adjectives, verbs and adverbs), before considering how these were used together or how they collocated with grammatical words, starting from the assumption that L2 learners' collocational use would be different from that of native speakers. Fan concentrated on language that was overused, underused or not used at all by these learners when compared to the British learners (Fan 2009:113). Findings revealed that, in general, the two groups used nouns and adjectives more frequently than verbs and intensifiers. Range of use of nouns, adjectives, verbs and intensifiers was on the whole higher in the British corpus – only three intensifiers (types) were found in the Hong Kong corpus, compared to the wide variety of types in the British corpus (Fan 2009:115). Hong Kong ESL learners tended to use more general and simple words instead of specific ones and seemed to have a preference for the amplifier (intensifier) *very* (see Granger 1998b for similar findings), where the British learners used the downtoner *quite*. Fan's findings suggested that the Hong Kong learners did not know as many words as their British counterparts, and that their collocational use was 'severely restricted' (Fan 2009:118) by their lack of knowledge of the vocabulary and grammar of English. They used fewer collocations than the NS students, and tended to overuse a smaller group of collocating words. She also found that the Hong Kong learners' L1 had a detrimental effect on their collocational use and that the Chinese learners' limited knowledge of English grammar and vocabulary restricted their use of collocations (Fan 2009:118). This

highlights the importance of vocabulary knowledge in collocational use and is particularly pertinent to the present study which explores, among other aspects, the link between vocabulary size or breadth (productive knowledge in this case) and use of MWUs, an aspect of depth of vocabulary knowledge.

This section has described various studies of collocation, using data from both NSs and NNSs, and the light they have cast on the difficulties learners have with these combinations. The findings of these studies underline the fact that prefabs, particularly collocations, should play a greater role in both ESL and EFL learning and teaching than they have hitherto, and they demonstrate that learners' phraseological skills are often severely limited. This is a point that informs the present study – the ability to use the right combination of words is something which non-native speakers of a language frequently find difficult, for many reasons but particularly because these combinations are often based on arbitrary or idiomatic rules. However, in order to design effective pedagogical tools, we need more empirical data on prefabricated language, including the type of MWUs explored in the present study, that is, 'good descriptions of learner use of prefabs' (Granger 1998b:159). Hopefully the present study will provide this type of evidence.

In the following section, the process of defining the particular type of MWU analysed in this study is discussed.

4.4 DELEXICAL MWUS

This section of the chapter deals with the specific type of collocation or MWU that is the focus of this thesis. One of the challenges in this study was arriving at a definition of this very specific type of combination; in the formulation of my definition, I was guided particularly by the work of Algeo (1995), Biber et al. (1999), Stubbs (2001), Langer (2004) and Nesselhauf (2003, 2004, 2005).

4.4.1 Defining delexical MWUs

In her study, Nesselhauf (2004, 2005) refers to these combinations as 'stretched verb constructions' (SVCs) (2005:211), that is, combinations of a delexical verb and eventive noun such as *make a decision*. She believes that 'what is special about these combinations is that the noun is derivationally related to a verb that is roughly synonymous with the whole combination: the meaning of *make an arrangement*, for example, largely corresponds to the meaning of *arrange'* (Nesselhauf 2005:20). As the verb carries very little meaning, or is semantically empty, it can be referred to as a 'delexical' (Nesselhauf 2004:109) or as a 'light' verb (Nesselhauf 2005:21). She remarks that 'the most restrictive definitions [of stretched verb constructions] only include combinations of one of the verbs *make, take, give*, and *have* with an indefinite article and an eventive noun that is identical in form to the verb with which the whole construction is roughly synonymous' (Nesselhauf 2005:21). This is in fact the definition I use for core delexical MWUs in

this study, and one which Algeo (1995) also uses, although he calls these combinations 'expanded predicates' and does not use the term 'delexical'. Nesselhauf (2005) makes the point that while these combinations have been the subject of a great deal of research, their relation to collocations is not often discussed, another reason for investigating them in my study. However, Nesselhauf makes it clear by her definition that, although some of these combinations may fall within her definition of restricted collocation, others may not. In many cases in her corpus these combinations are regarded as deviations, used where single-word verbs would have been more appropriate; she found that in many instances the learners in her study used these SVCs where 'an "unstretched", i.e. a derivationally related verb' was regarded as more idiomatic. She provides as an example *give me relaxation*, instead of the more appropriate *relax me* (Nesselhauf 2005:112).

The MWUs investigated in this study are combinations featuring one of three high-frequency verbs, *HAVE*, *MAKE* and *TAKE,* used delexically with an 'eventive object' (Algeo 1995:204) or 'deverbal noun' (Live 1973:35). Stein (1991:5) describes the eventive object (the noun in the combination, that is) as 'semantically an extension of the verb … [that] bears the major part of the meaning'. Live (1973:32–3) explains 'deverbal noun' as the word in such combinations that 'bears the lexical load of the phrase. This word has the formal attributes of a noun […] and has been referred to as a deverbal (or as a zero-derivation noun)'. These are combinations such as *have a look*, *make a claim, take a walk*. In these combinations, the verb is used delexically; that is, it carries little semantic content of its own while the noun carries the weight of the meaning. In examples such as the ones above, the meaning appears to be mostly carried by the noun, and many such phrases mean the same or almost the same as corresponding single-word verbs. For instance, the examples above could be replaced by the verbs *look, claim* and *walk*. Such delexical combinations are very common in speech and writing. Stubbs (2001:32–33) made a search for the lemma *TAKE* in a corpus of over two million words and while he found 400 examples, this lemma was used in its literal sense of 'grasp with the hand' in only about 10% of these occurrences. What he did find, however, was that by far the most common use of the verb *TAKE* was in delexical combinations such as *take a deep breath, take a photograph, take a decision* (Stubbs 2001:32). He emphasises that delexicalisation is common in English, and quotes Sinclair (1992, cited in Stubbs 2001:33), who claims that the meaning of words when they are chosen together is different from their independent meaning; they are at least partly delexicalised.

Algeo (1995), in his detailed study of this aspect of phraseology, agrees with Stein (1991) and others researching in this field that there is as yet 'no generally acceptable term for this construction', that is, for 'the grammatical/lexical construction such as *have a look'* (1995:203). He calls it an 'expanded predicate'; this is 'an idiomatic verb-object construction in which the verb (e.g. *do*, *give*, *have*, *make* or *take*) is semantically general and the object is semantically specific (such as *somersault*, *nod*, *rest*, *promise*, or

walk)' (Algeo 1995:203–204). These constructions have also been called 'light verbs' in a 'phrasal' pattern (Live 1973:32), 'periphrastic verbal constructions' (Wierzbicka 1982), 'phrasal verb types' (Stein 1991), 'stretched verb constructions' (Nesselhauf 2004, 2005:20–21) and 'support' verb constructions (Langer 2004).

Langer (2004:3) defines delexical combinations as 'combinations of predicative nouns and semantically weak/reduced verbs, where the noun categorizes semantically, and the verb syntactically'. He explains a predicative noun (called an eventive noun in the present study) as one which signifies an action or state. 'The noun has an argument structure, i.e. it subcategorizes at least one semantic participant' (Langer 2004:6). The semantics (meaning) of the support verb 'is either void or reduced to a small set of semantic features that are relevant to very large subclasses of verbs' (Langer 2004:6). In his paper, Langer discusses the difficulty of finding these constructions by tagging, a point which I take as added support for taking an essentially corpus-driven approach to these combinations myself.

In the opinion of Biber et al. (1999), these combinations are restricted MWUs which occur habitually (see Laufer and Waldman 2011:648) and are relatively transparent in meaning or 'semi-compositional' (Erman and Warren 2000; Langer 2004:6). The combinations Biber et al. (1999) refer to, like those investigated in this study, differ thus from free combinations, where individual words can easily be replaced by other, grammatically appropriate words; although most allow for some replacements, these restrictions are sometimes seemingly arbitrary. Biber et al. (1999:1026) explain the phenomenon thus: *have*, *take* and *make,* for instance, combine with 'a following noun phrase to form relatively idiomatic expressions' which 'form a cline of idiomaticity' between expressions which are clearly idiomatic, such as *have a look*, *make a killing*, *take time*, and those which 'retain the core meaning of these verbs', as in '*we have an extra one*, *he made a sandwich*, *you can take a snack in your pocket*'. In the middle are a host of 'relatively idiomatic' phrases 'such as *have a chance*, *make a statement*, *take a walk'* (p. 1027). Biber et al. (1999:1027) also note that although the 'core meaning' of the individual words is retained, all these types of expression do tend to take on a more idiomatic meaning. These scholars, like Stubbs (2001), also note that most of these expressions can be replaced by a single verb.

This last aspect is particularly critical to my own study, in that these combinations are regarded as *core* (Algeo 1995) delexical MWUs only when the single-word verb and the noun are identical, both in form and in meaning, e.g. *have a sleep* = *sleep*, *make a mess* = *mess*, *take a walk* = *walk*. If the form of the noun is different in any way, for instance because it is plural, or because it is a derivation formed by affixation, or in the passive voice, or if another single-word verb is required to provide the same meaning as the V + N combination, the combination is classed in my study as a pseudo delexical MWU (Algeo 1995), for example, *have the disgrace*, *make love*, *take a decision*. Similarly, if a definite article is used instead of the indefinite

article, or no article, as in *have control,* either because that is the grammatically acceptable usage or because it has been used in error instead of the indefinite article, the combination is classed as a pseudo delexical MWU in my analysis because it is in most cases impossible to judge whether this is simply a slip of the pen or the way in which the student habitually uses the combination. As far as errors as opposed to mistakes are concerned, I investigated the corpus and extracted all delexical combinations. Thus, if such an MWU was not a core delexical MWU it was in almost all other cases a pseudo delexical MWU; but whether this came about through a mistake or through misunderstanding the rules of the combination (an error) or by the restrictions of usage was not always clear. For this reason I did not take the analysis any further than simply distinguishing core from pseudo delexical MWUs. As far as restriction of these MWUs is concerned, Howarth's (1998b:169–170) category of restricted collocations (Levels 2 to 4) applies here (see §4.2.2).

According to this definition, then, these MWUs are collocations made up of monotransitive[17] verb patterns (verb + noun) which occur on Howarth's (1998a, b) continuum, or on Erman and Warren's (2000) cline, somewhere between free combinations and idioms. Some of the expressions, such as *take care*, are regarded by Biber et al. (1999) as pure idioms, others simply as idiomatic. These combinations are more transparent than idioms, but this does not mean that they are always compositional (Erman and Warren 2000:54). The fact that they are not always compositional is what makes them so difficult for learners to master.

The following two sections focus on studies of MWUs featuring high-frequency verbs and the challenges these pose for learners.

4.4.2 Studies investigating high-frequency verb-noun combinations

This section deals with studies which have investigated MWUs comprising combinations of high-frequency verbs and nouns. As already indicated in Chapter 3, two seemingly contradictory observations have emerged about high-frequency verbs: these verbs tend to be overused by learners: 'core words – learnt early, widely usable, and above all safe (*because* they do not show up as errors) are hugely overused, even among learners sufficiently advanced to have been weaned off them' (Hasselgren 1994:250; for similar findings, see Altenberg and Granger 2001; Kaszubski 2000; and Ringbom 1998). However, learners also experience particular difficulty with these words, especially when they are used in multiword combinations that are restricted or semi-restricted (see §4.4.3). In his 'underuse hypothesis' Sinclair explains that many learners avoid these high-frequency verbs, especially where they are part of idiomatic phrases, making use instead of 'larger, rarer or clumsier words which make their language sound stilted and awkward' (Sinclair 1991:79).

[17] A monotransitive verb takes one object.

Hasselgren's study (1994:242) found, contrary to Sinclair's (1991) underuse theory, that high-frequency words such as *get* (what she calls 'core' words, high in frequency, neutral in meaning and collocating widely) were significantly *over*used, resulting in mismatches and dissonances which are primarily collocational. Often these utterances are not wrong or inappropriate in meaning but in usage. She explains this as 'an item may be felt to be inappropriate simply through disharmony with the words around it without being deviant in meaning or style' (Hasselgren 1994:243).

Based on Biskup's (1992) finding that L2 learners had most difficulty with collocations involving verbs and objects, Hasselgren (1994:255) posited that if learners are faced with a 'difficult lexical decision' their tendency will be to use a core item. Using a task she designed herself, Hasselgren investigated the use of transitive core verbs by advanced learners of English and native speakers. She provided subjects with the direct objects *treatment*, *identity*, *reputation* and *sympathy* and asked them to complete the sentence by filling in a verb. On examination of the verbs the respondents provided, she defined the following as core words: *give, get, take, have, know, keep, tell* and *make*. She found that although both groups preferred to use these core words rather than more native-like collocations, such as *administer treatment*, for example, learners used them significantly more than native speakers, underlining once again the overuse of familiar items by learners (Hasselgren 1994:255–256). She also found that the types of wrong word choices learners made tended to lead to wrongness of collocation or style rather than of meaning (1994:256); learners seemed to have little intuition when it came to native speaker patterns. This, Hasselgren (1994:256–257) felt, provides a convincing argument for teaching language in chunks rather than focusing on words as isolated items. This also supports the argument for encouraging more extensive reading which would provide learners with greater exposure to written language, a point which is expressed elsewhere in this thesis and one which is particularly relevant to learners in South Africa. Like Nesselhauf (2004, 2005) and Granger (1998b), Hasselgren found that the overuse of wrong core words and synonyms was clearly influenced by the L1. However, Hasselgren found that 'at least 70% of these synonyms are selected in an environment of divergence; in other words, both the wrong synonym and the correct alternative are translations of the L1 source item' (1994:250). These findings might go some way to explaining why teachers are more aware of some lexical problems than others. It may also explain why they are less aware of the problems caused by the wrong selection of synonyms or by errors in association (1994:249–250). What Hasselgren concludes, and what I hope my study will show too, is that words do not have to be 'wrong' to be different or out of place, unidiomatic or unnative-like. Such words are what Hasselgren calls 'lexical teddy bears' (1994:251) – words that are familiar because there are direct translations in the L1, or because they have been learnt early on in the process of language learning and learners cling to them because they are seldom marked 'wrong'.

Ringbom (1998) used seven western European learner corpora from the ICLE database and found that most of the top 100 most frequent words were function words with grammatical meaning. His findings revealed that learners tend to overuse auxiliary verbs such as *be*, *have*, *do* and *can*. NSs used fewer of these high-frequency verbs but their usage also revealed a different distribution of these words from NNSs. Only four verbs (*make*, *use*, *believe* and *feel*) had a frequency of 10 or more per 10 000 words in the NS corpus, with *make* the most frequent of these. NNSs also made great use of *make,* but they tended to use verbs such as *think* and *get* even more than other high-frequency verbs, and their use of high-frequency verbs in general was on the whole much higher than that of the NSs (1998:43). The overuse of the verb *get* was particularly evident in the ICLE corpus and underlined these students' lack of collocational competence; they were inclined to use words in contexts where NSs would not choose them (see Hasselgren 1994).

Ringbom (1998:50), like Fan (2009), remarks on the limited vocabulary of advanced learners: although frequency lists revealed that a few common words were underused by learners, a lot more were overused, regardless of the learners' L1. This overuse is one of the main reasons for learner language being criticised as dull, stereotypical and vague, in Ringbom's view. In his opinion, the non-native features of the ICLE corpus are 'less due to errors than to an insufficient and imprecise, though not necessarily erroneous, use of the resources available in English' (1998:51).

Hasselgren's (1994) and Ringbom's (1998) findings are particularly relevant to my study – it has become clear that it is not always the meaning of a word that is problematic for learners, but rather how this word is used, particularly those simple, high-frequency words that are so taken for granted by both learners and teachers. This can make learners' expression, spoken or written, sound unidiomatic and lacking in fluency.

Altenberg and Granger's (2001) study, and those by Howarth (1998a, b), Kaszubski (2000), Lee and Chen (2009), Liu and Shaw (2001), Nesselhauf (2003, 2004, 2005) and Wang and Shaw (2008), show that this difficulty is a universal one, although error types do seem to be partly L1 specific. Although these 'core' verbs (Kaszubski 2000; Paquot 2010) appear to be monosemic, in actual fact they are polysemic because 'their meanings are obscured or confused with contextual, inferential meanings' (Carter 1987, cited in Liu and Shaw 2001:173; Altenberg and Granger 2001). These words also attract more collocates because of their polysemy (Kaszubski 2000:15).

By taking a contrastive corpus analysis approach to the investigation of the use of the word *make* in two main corpora – the CSLE (Chinese-speaking Learners of English) and the NSE (Native Speakers of English) corpora – Liu and Shaw (2001:174) identified the 'idiosyncrasy' of EFL/ESL learners' word usage in an effort to suggest strategies for vocabulary teaching. They support the analysis of these words in a study such as the one described in this thesis because they are not usually the focus of ESL/EFL vocabulary instruction.

They found that learners of English used *make* (and its various inflections and derived word forms) far more frequently than native speakers, regardless of the category of text considered (Liu and Shaw 2001:176). These scholars question whether this overuse is unique to Chinese learners or a feature of all NNS learners writing at this level. This is an aspect which the present study will also address: will the writers in the student corpus show the same tendency to overuse high-frequency, more familiar lexical items to convey their meaning?

Liu and Shaw (2001) do not specify a delexical verb category when looking at *make* – they distinguish several categories according to the grammatical properties of the word, but they call this particular usage 'complex verbal phrases' (*make* + *n* + *prep*), such as *make use of*, *make fun of*, *make a mockery of*, which they include with their definition of phrasal verbs (which they describe as *make* followed by an obligatory particle, an adverb, a preposition or both, such as *make out*, *make off with* [Liu and Shaw 2001:177]). The present study takes something from Liu and Shaw (2001), but develops the category of delexical combinations further and focuses on that. But these scholars repeat the point made by so many researchers mentioned in this thesis: that verbs like *make*, *take* and *have* tend to co-occur with other words to form prefabricated combinations, the components of which are 'mutually expectant' (Liu and Shaw 2001:183). They believe that this is a 'strength which contributes to the versatility' of these verbs (Liu and Shaw 2001:183). Thus they would regard the MWUs in this study as restricted or semi-restricted combinations; there is a degree of freedom in what words may be combined, such as *make a decision*/*mistake*/*speech*, but there is a restriction on the number of items that can be combined with these verbs (2001:184). It is thus idiomatic and acceptable to native speakers to say *make a decision* or *a choice* or *a selection*, but not *make an option* or *make a determination*. Liu and Shaw (2001) found that the senses of *make* were more specific and more varied in idioms and phrasal verbs than in freer combinations. They also make an important observation that the 'closer the sense of *make* is to its core or root meaning, the more objects it can combine with, and the more metaphorical, the more fixed it is the less freedom it allows' (Liu and Shaw 2001:183). Their study thus deals with high-frequency vocabulary but also provides insights into delexical verbs, which the studies discussed in the next section have taken further.

In a comparison of the collocational errors of Swedish and Chinese university students learning English, Wang and Shaw (2008) noted that advanced learners, in particular, experienced difficulties with collocations formed with what they refer to as 'frequent, high-utility dynamic verbs' (2008:203), citing Ringbom (1998), who looked at advanced learners' use of the verb *get.* Both Wang and Shaw's (2008) and Ringbom's (1998) studies revealed a lack of collocational competence among learners from various language backgrounds when using this verb.

It is clear from this section that certain high-frequency verbs can be a source of difficulty for learners. As Partington (1998) puts it: 'what is formally possible' in a language 'corresponds to Chomskian competence' and is not governed by context, while what is feasible, appropriate and what is actually performed 'are context-related, that is, they are knowledge systems which permit language users to make certain decisions about language use depending on the situation they find themselves in and the requirements they have' (Partington 1998:18). In other words, while grammar governs what could occur in language, this will be limited by what is psychologically feasible, sociolinguistically appropriate and actually frequent, but also typical, among speakers of the language (Stubbs 2001:61).

In the following section, empirical studies which have focused specifically on delexical uses of high-frequency verbs are discussed, as it is this type of MWU which is the focus of this study.

4.4.3 High-frequency verbs used delexically

Stubbs (2001) believes that such MWUs as those investigated in this study, though they may not be idioms, are idiomatic, in that they are used in a particular way, sometimes arbitrarily, by native speakers. That is, they are what he calls transparent (see Erman and Warren 2000:54): they are more transparent than idioms, but this does not mean that they are always *compositional* – they are not difficult for learners to decode (read and understand) but they cause difficulties when it comes to encoding them – because the learner simply has to know the conventional way of saying these things (Fan 2009:111; Farghal and Obiedat 1995; Martinez and Schmitt 2012; Nesselhauf 2003:224).

That the common verbs featured in these MWUs are often 'error-prone' (Altenberg and Granger 2001:179; Stubbs 1991; Yan 2006) in learner language, both written and spoken (Laufer and Waldman 2011; Howarth 1998a, b; Lennon 1996) is well documented. For instance, using a computerised corpus of EFL writing made up of two samples from the ICLE database, one of writing by advanced French-speaking learners of English and the second, of a comparable size, made up of writing by Swedish learners, Altenberg and Granger (2001) focused on the grammatical and lexical patterning of the verb *make* in students' writing. They were particularly interested in whether learners overused or underused high-frequency verbs like *make*. They also wished to investigate whether high-frequency verbs were particularly likely to cause errors, and the part played by L1 transfer in the misuse of these verbs (Altenberg and Granger 2001:178). They note the following characteristics of high-frequency verbs which might make them problematic for learners: they express basic meanings; they have equivalents in most languages; they are highly polysemic because they tend to be used delexically (an important point in the light of the present study); but they also have language-specific, specialised meanings, especially in collocational and idiomatic uses.

Altenberg and Granger (2001) found that the Swedish and French learners underused the verb *make* in delexical structures, while conversely overusing the same verb when it was used causatively (e.g. *to make somebody believe something*). They examined collocates of *make* that occurred at least twice in the corpora; one of the main differences between the NS corpus and the two learner corpora was in the frequency of 'speech' or 'verbal communication' collocates – *argument, claim, point, statement, case*. Almost a third of NS tokens belonged to this category of speech nouns, while only 9 to 13% of learner instances did. The learner corpora also confirmed Sinclair's 'underuse hypothesis' (Sinclair 1991:79), revealing that learners may simultaneously overuse a high-frequency verb and underuse its delexical structures.

NNSs also misused these delexical structures. This was in fact the category that accounted for the majority of learner errors with *make* in the corpus, some of which could be explained in terms of interlingual interference. These errors also illustrated the difficult choice between *make* and *do* facing EFL learners. Results of this study show that, even at an advanced level of proficiency, EFL learners have great difficulty with high-frequency verbs such as *make* (Altenberg and Granger 2001:189). Altenberg and Granger's (2001) findings are particularly relevant to the present study, as they support the very premise of the choice of MWUs for investigation in this study.

More recently, Laufer and Waldman (2011) also noted the importance of these delexicalised verbs and their collocations in their study of verb-noun collocations in the writing of three groups of Hebrew-speaking second language learners at different proficiency levels. They found that the NNSs used fewer collocations overall, using significantly fewer of these combinations than the NSs, revealing that collocation poses difficulties even for advanced learners (2011:664–665). As in Altenberg and Granger's (2001) study, Laufer and Waldman (2011:664) found that, compared to native speakers, learners tended to *under*use collocations containing what they refer to as 'core' verbs (*be, have, make*) or those with particular amplifiers (*very, completely, highly, strongly*).

Laufer and Waldman (2011) stress the importance of MWUs, which are 'currently viewed as a necessary component of second-language (L2) lexical competence in addition to the knowledge of single words' (Laufer and Waldman 2011:648), and they underline the now commonly accepted view that knowledge of MWUs improves both the quality and the fluency of language, both spoken and written. They mention studies which have found that learners who have control of the idiomatic dimension of language sound both proficient and fluent; these constructions are an aid to fluency and idiomaticity. Such control also sets advanced learners apart from intermediate ones (Thornbury 2002, cited in Laufer and Waldman 2011:648). Gledhill (2000, cited in Laufer and Waldman 2011:648) believes that 'it is impossible for a writer to be fluent without a thorough knowledge of the phraseology of the particular field he or she is writing in'.

Laufer and Waldman (2011) note, like Howarth (1998a), that this has much to do with the fact that the procedural vocabulary[18] of academic disciplines consists in large part of predicate structures such as *make a claim*, *reach a conclusion*, *adopt an approach* and *set out criteria*. Such structures are often made up of high-frequency verbs and eventive nouns. If learners are unaware of these conventions, the comprehensibility of their speech or writing may be affected: they may not be able to express *what* they know about a subject, either precisely or concisely.

Laufer and Waldman (2011:651) observe that several corpus studies have investigated MWUs comprising collocations of high-frequency verbs and nouns through a comparison of learner data with native speaker corpora (Bahns and Eldaw 1993; Farghal and Obiedat 1995; Hasselgren 1994). For instance, Kaszubski (2000), looking at the high-frequency verbs often used in delexical structures (in this case *be, have, make, take, do, get*) and their various combinations in corpora of Polish, Spanish and French learners of English, found that delexical collocations were underused.

As noted above, Nesselhauf (2003, 2004, 2005) separates collocations from stretched verb constructions (SVC), while I do not (§4.4.1). She notes that SVCs, or delexical MWUs, have been found to be particularly difficult for language learners (2005:211), mentioning Altenberg and Granger (2001) as a study which has revealed this finding. In her study, she found that these combinations made up over 20% of all the collocations she identified in her German learner corpus (Nesselhauf 2005:211). Students used *have*, *make*, *do*, *give* and *take* most often (in that order), with a few occurrences of *find*, *come to*, *live*, *put* and *pay*. Almost a quarter of these SVCs were judged as unacceptable. When those which were questionable were added, 43% of all SVCs were deviant. She found that students were more inclined to make errors in combinations where the verb takes a relatively wide range of nouns, than where the number of nouns a verb can take is more restricted (what she terms RC1 restricted collocations), such as *pay attention*, *take a picture* (Nesselhauf 2003:233). This may be because such RC1 combinations are 'more often acquired and produced as wholes' whereas the less restricted collocations allow for more creative combinations, and thus more potential errors (2003:233).

This finding suggests that the learners in Nesselhauf's study (2005:212), contrary to other research findings, did not find these combinations as difficult as collocations which were not SVCs – and she mentions Howarth's (1998b) study which had similar findings. Thus, SVCs with light verbs, especially those with high-frequency light verbs, such as *live a life – live, give solutions to* – solve, and *give them respect – respect*, were not in the main produced incorrectly when they were used by learners – but because learners used these structures often and because they often contain high-frequency verbs, they got them wrong in the

[18] Procedural knowledge is 'knowledge of how to perform an activity' (Richards and Schmidt 2002), so procedural vocabulary would be that which expresses the *how to* aspect of particular disciplines.

'absolute sense' (Nesselhauf 2005:212). In other words, her argument is that because learners tend to use these combinations frequently, their mistakes in using these high-frequency verbs are seen often, leading to the judgement that learners find them difficult.

In a study of spoken English, Shirato and Stapleton (2007) compared a small corpus of spoken language from adult Japanese NNSs of English with an established NS corpus (conversational components of the BNC). The authors found that Japanese NNSs differed particularly in their underuse of, among other features, delexical verbs. Although much of the emphasis in corpus linguistics to date has been on written texts, Shirato and Stapleton (2007) showed that NNSs used delexical verbs relatively infrequently in their speech when compared to NSs, seemingly reluctant to use expressions such as *get done* and *get locked* in, for example, and unaware that such delexical words may be more appropriate in spoken language than more formal single-item verbs (Shirato and Stapleton 2007:407). These findings are supportive of those of Altenberg and Granger (2001) and Laufer and Waldman (2011) and show that even advanced EFL learners can have great difficulty with high-frequency verbs such as *have, make* and *take*.

Similarly, in their study Lee and Chen (2009) also took a multiple-comparison approach to an investigation of words and phrases which were overused and underused by learners. They compiled 'three complementary types of corpora' (2009:151). The first corpus, the Chinese Academic Written English (CAWE), consisted of dissertations by Chinese undergraduates majoring in English linguistics or applied linguistics. In addition, they used the Expert Journal Articles (EXJA) corpus, made up of articles from various linguistics and applied linguistics journals. Their third corpus was a section of the British Academic Written English (BAWE) corpus, comprising high scoring student essays written by native English speakers, which they called the sub-corpus BAWE-L. This last corpus was used to validate results from the comparisons between the CAWE and EXJA corpora; they also felt that this 'would represent, in terms of academic lexico-grammatical features, a kind of half-way house between EFL learner texts and published journal articles written by experts, and therefore something of a more realistic yardstick' (2009:152).

Using a keyword analysis and a 'bottom-up, corpus-driven methodology', Lee and Chen (2009) used frequency information to identify salient patterns for further investigation. This technique allowed them to identify both positive keywords, those that were overused when compared to the reference corpus, and negative keywords or those that were significantly less frequent or 'underused' (2009:152). They found that most of the overused words and phrases were closed-class or function words such as *can*, *the*, *some*, and very high-frequency common words such as *make, besides, get, help*. They observe that these words are not focused on in EAP, where the stress tends to be on 'academic words' in lists such as the Academic Word List (Coxhead 2000). They then conducted qualitative investigations of concordances and collocations of the key items, focusing on five key items and their associated clusters. Like Altenberg and Granger (2001),

Lee and Chen (2009) found that the high-frequency verb *make* was used much more frequently by Chinese learners in what they term 'light verb' or 'delexical' constructions than by writers in the EXJA and BAWE-L corpora. However, many of these expressions were in actual fact unidiomatic or non-nativelike; these scholars observed that the function words and so-called 'simple' words in their lists occurred in 'recurrent problematic patterns rather than being randomly used' (Lee and Chen 2009:154). They also found that *make* was often used by learners as a substitute for other causative verbs in usages which were 'unnatural' or 'informal or colloquial' (2009:156). Also noticeable in students' writing were examples of what Lee and Chen (2009:158) call 'marked' usages, both pragmatic-stylistic and collocational, which though not actually wrong, were awkward and unidiomatic. They cite the example of *make a conclusion,* which was used by the NNSs; the NSs used *conclusion* with other verbs such as *draw, reach* and *arrive* (2009:155). What they noted particularly was that Chinese learners need training in the 'subtle facets' (Lee and Chen 2009:159) of meaning of many of the academic expressions that they had been taught to use in their writing.

This study by Lee and Chen (2009) supports the value of using learner corpora to identify difficulties in academic writing and attempting to rectify them. Chinese learners, like many South African learners, have little exposure to expert writing other than in their textbooks. Much of the time teachers of English in these contexts are not mother tongue speakers of the language and often lack the native speaker's intuition about 'subtle, pragmatic and stylistic issues' (Lee and Chen 2009:160). What is most relevant about their study, however, is the finding that many of the patterns they identified as problematic contained common high-frequency words and function words that occur frequently in collocational patterns which we would do well to teach our students. In other words, academic vocabulary alone should not be the entire focus of our instruction.

This section has provided a brief overview of some studies which have reflected the difficulties learners experience with the reception and use of MWUs containing delexical verbs. It therefore seems reasonable to examine the behaviour of this type of collocation in both expert and novice writing, and in this way to identify some of the difficulties that student writers may experience in the use and identification of these constructions. This should, it is hoped, lead to greater insight into the problems among South African university students in particular and may assist in developing strategies to assist learners in general in recognising and using these combinations.

4.5 CONCLUSION

In this chapter I have discussed formulaic language in general and delexical MWUs in particular. The findings of these studies make it clear that these word combinations pose challenges for learners. These units of language, in their many guises, are regarded as an essential component of lexical competence,

together with the knowledge of individual words (Laufer and Waldman 2011:648), and as particularly crucial for second language learners to master. Their use improves the fluency and quality of both spoken and written language (Fan 2009; Nesselhauf 2004; Pawley and Syder 1983; Shirato and Stapleton 2007; Wray 2002). Howarth (1998a:34) believes that phraseological competence is vital to successful style; clarity, precision and a lack of ambiguity are crucial to academic register, for instance. A student must develop an awareness of the arbitrary restrictions that exist on collocation (Howarth 1998b:162), particularly when writing in an academic context. Learners need to understand that such restricted collocations make up a significant part of a typical native speaker's speech and writing. One of the greatest challenges facing the non-native writer and speaker is knowing which range of collocational options are restricted and which are free (Howarth 1998a:36): 'this involves distinguishing between what is semantic and thus generalisable and what is collocational and thus highly specific' (Howarth 1998a:42).

The current study fits into this area of linguistic study, in that it investigates the production by a group of student writers, both NSs and NNSs, of delexical verb + noun collocations, and their ability to distinguish between the 'semantic' and the 'collocational', that is, their understanding of the sometimes arbitrary restrictions that idiomatic English places on collocations, in an effort to establish whether the restrictedness or semi-restrictedness of these MWUs and their non-compositional nature makes for greater difficulties for students with less adequate vocabularies. This is an aspect of vocabulary studies which has been addressed relatively infrequently by researchers and I hope that my study will add value to the body of knowledge on MWUs, particularly in the context of student academic writing. I hope in this study to shed more light on the difficulties experienced by the students in my sample in the use of MWUs, and in this way to add to what researchers in the field have revealed about this linguistic phenomenon.

Chapter 5

Methodology

5.1 INTRODUCTION

The focus of this chapter is the description of the methods and processes followed in this study. The chapter is organised according to the three phases of the study; these are briefly described together with the research questions driving each phase. Thereafter, the research design is explained, followed by issues of validity and reliability, that is, issues related to methodological rigour. The three phases of the study are then described in detail. Because Phases 1 and 3 involve similar quantitative methodologies, they are described first (§5.7 and §5.8), followed by Phase 2 (§5.9), which involves both quantitative and qualitative approaches. Finally, justification for the techniques used in this study is provided and ethical considerations are discussed.

5.2 RESEARCH QUESTIONS AND PHASES

As indicated in Chapter 1 (§1.5), this study was designed to provide answers to six main research questions, each aligned with a research phase:

Research Question 1 (Phase 1)
What is the size of the productive vocabulary of undergraduate students?

Research Question 1 was broken down into four sub-questions:

Research Question 1.1: What is the size of the productive vocabulary of students at each level of the VLT, within and across two courses?

Research Question 1.2: Are there significant differences in the size of the productive vocabulary of male and female students?

Research Question 1.3: Are there significant differences in the size of the productive vocabulary of students from different age groups?

Research Question 1.4: Are there significant differences in the size of the productive vocabulary of students from different language backgrounds?

Research Question 2 (Phase 1)

What is the relationship between the size of the students' productive vocabulary and their academic performance?

Research Question 3 (Phase 2)

How does the distribution of the functions of the three selected verbs compare within and across the Expert and Student corpora?

Research Question 3.1: How does the distribution of the functions of the three selected verbs in the Expert Literature corpus compare with the distribution in the Expert Law corpus?

Research Question 3.2: How does the distribution of the functions of the three selected verbs in the Student Literature corpus compare with the distribution in the Student Law corpus?

Research Question 3.3a: How does the distribution of the functions of the three selected verbs in the Expert Literature corpus compare with the distribution in the Student Literature corpus?

Research Question 3.3b: How does the distribution of the functions of the three selected verbs in the Expert Law corpus compare with the distribution in the Student Law corpus?

The second part of the analysis involved a closer focus on the MWUs containing the selected delexical verbs, an analysis of these MWUs for deviance and a focus on the errors within the MWUs that gave rise to this deviance. These concerns are addressed in Research Question 4:

Research Question 4 (Phase 2)

How does students' use (in terms of frequency and deviance) of selected MWUs compare with the use of these MWUs by expert writers within and across courses?

This question was split into three sub-questions:

Research Question 4.1: How does the use (frequency) of MWUs in the Student Literature corpus compare with their use in the Expert Literature corpus?

Research Question 4.2: How does the use (frequency) of MWUs in the Student Law corpus compare with their use in the Expert Law corpus?

Research Question 4.3: How does the use (frequency and deviance) of MWUs in the Student Literature corpus compare with their use in the Student Law corpus?

The 'Expert' corpus is named thus because even though it is made up of writing by academics from several language backgrounds, this writing was put through stringent editing, peer review and critical reading stages before being published and can therefore be regarded as expert. Thus, in this study the actual language background of the individual authors is irrelevant; the writing in the Expert corpus is regarded as a model of the sort of academic writing to which undergraduates at this level should aspire. Because there were no deviations in the Expert corpus, deviations were only compared across the two Student corpora.

The MWUs referred to in this study are described in greater detail later in this chapter (§5.9.2). The study attempts to highlight areas of difficulty in this sample of student writing, and the types of deviation and error which characterise such writers' production of these structures, and to identify differences between student and expert writing. The differences in MWUs between the texts written by students studying Literature and those in the Law course were also investigated; the Expert and Student corpora were both split into two sub-corpora, one of Literature texts and one of Law, for this purpose.

In the last phase, Phase 3, two research questions were addressed:

Research Question 5 (Phase 3)
What is the relationship between the size of students' productive vocabulary and their production of selected MWUs, within and across courses?

Research Question 6 (Phase 3)
What is the relationship between students' production of selected MWUs and their academic performance, within and across courses?

5.3 RESEARCH DESIGN

Research can be quantitative or qualitative, or a combination of the two. In quantitative research, data collection results in mostly numerical data which are analysed using statistical methods; in qualitative research, on the other hand, the data are usually non-numerical and are analysed by non-statistical methods. This might include the analysis of transcriptions of interview responses, for example. A third type of study may be mixed, employing both methods; mixed method research will combine aspects of quantitative and qualitative research in the data collection process or in the analysis of the data (Dörnyei 2007:24).

Quantitative and qualitative studies differ in another fundamental way. Quantitative studies, though they may include inductive, exploratory research and descriptive statistics as this study has done, tend to be

deductive in nature, and hypothesis driven, in that the researcher will develop a hypothesis based on known theory about the relationship between two or more variables. The empirical investigation then attempts to support or disconfirm this hypothesis. Qualitative research, on the other hand, is essentially inductive, in that theory is derived from the results of the study. It is thus exploratory and data driven.

Research aims and questions determine the research design and so quantitative and qualitative studies will naturally differ in the way in which the aim of the study is specified and the way in which this is broken down into specific research questions. Quantitative research often explores relationships between variables and questions are thus specific and require specific methodological procedures (Dörnyei 2007:74). Qualitative research studies, on the other hand, are often what Dörnyei (2007:74) calls 'emergent' in nature, meaning that research questions tend to be less specific than those in quantitative studies, and qualitative research often sets out to explore a particular phenomenon or idea in order to gain a better understanding of it and even to develop a new theory.

This study is predominantly quantitative in design and method, although there is a strong element of qualitative corpus linguistic methodology in Phase 2. Phase 1 is quantitative in that it includes the measurement of students' productive vocabulary size and the relationship this has to factors such as academic courses, gender, age and language background, and to their academic performance, measured by their examination scores. Phase 3 is also quantitative in that it investigates the relationship between the student writers' production of MWUs and their productive vocabulary size, as measured by the VLT, and their academic performance, as measured by their examination scores.

Phase 2 is corpus-driven although it does include an element of corpus-based methodology (see Biber's definitions and explanations, §2.4.3) – 'lexical collocations of a target word, for example, are discovered through corpus analysis [corpus-driven], but a preliminary stage exists where the linguist chooses interesting words [corpus-based]' (Biber 2009:276). The corpora have not been tagged and the analysis of the concordance lines is inductive rather than deductive: the focus is on the patterns and clusters that emerge from a concordance of the corpora, although the particular constructs are identified at the outset, namely, the high-frequency verbs *HAVE*, *MAKE* and *TAKE*[19], and the multiword combinations featuring these verbs used delexically. In this study, the high-frequency verbs that occurred most commonly in both corpora were, in frequency order, *HAVE DO*, *MAKE, TAKE, GET* and *GIVE.* Although frequency for *DO*, *was* high, I chose to include *MAKE* and *TAKE* instead of *DO*, because *MAKE* and *TAKE* are more productive of the type of MWU in which I was interested (cf Biber et al. 1999). All three of the verbs chosen for the study are also highly polysemous (see §1.3.2).

[19] From this point on, the capitalised forms, *HAVE*, *MAKE* and *TAKE*, indicate the word or lemma, while the words in lower case, such as *have*, *has*, *had*, indicate those specific word-forms of the lemma.

Such combinations make up propositional statements in academic writing and I assumed, prompted by findings by scholars such as Altenberg and Granger (2001), Howarth (1998a, b), Nesselhauf (2003, 2004, 2005), Langer (2004) and Kaszubski (2000) among others, that because of their high frequency of occurrence in general and their documented problematic nature for learners they would be worth exploring more thoroughly. Biber (2009) observes that there are no hard and fast rules and it is quite acceptable to integrate corpus-based and corpus-driven approaches in this manner. What he calls a hybrid approach is generally regarded as acceptable. Like Biber (2009), I do not consider that one approach is superior to another; I believe that the fact that I have combined two approaches that can provide 'radically different perspectives on language structure and use' (Biber 2009:279) and the different methods they require has strengthened this study.

The analysis in Phase 2 required quantitative methods although much of this phase also involved qualitative analysis. The generation of wordlists for a corpus reveals how many times each word occurs in that corpus – but the analysis of a concordance for a particular word reveals not only how many times that word is used in the corpus, but its context each time it is used, allowing for a qualitative investigation of each use of the word. In this study I searched the concordance lines manually for patterns and specific combinations, analysing each occurrence and placing it in a particular category. It is possible to use an automatic tagger to search a corpus, depending on the categories one is interested in, but in this case there was no tagger available which would have been able to isolate all the occurrences of the three verbs I was investigating in the detail I required. I focused on fine functional distinctions that apply to the three selected verbs, such as the identification of delexical uses of these verbs, and the tagsets I am aware of – such as the Penn Treebank[20] – simply do not subcategorise so finely. I thus had no option but to do all categorisation manually. The identification and description of deviant MWUs was a qualitative process: although these categories were, of course, grammatically defined, the 'stylistically marked' uses of these MWUs (cf. Lee and Chen 2009:151) were qualitatively investigated and analysed:

> Although corpus linguistics is often seen as a quantitative form of analysis, in fact human input is required at almost every stage, from corpus building (deciding what should go in the corpus) to corpus analysis (what research questions should be asked, what should be looked for, what analytical procedures should be carried out, how results can be interpreted) (Baker 2010:109).

Once all the selected MWUs had been identified they were quantified and analysed statistically. Baker notes that 'in using large amounts of naturally occurring data, corpus analysis offers a high degree of reliability and validity to linguistic research' (2010:111), characteristics that are typically associated with quantitative research.

[20] Available online at: http://www.cis.upenn.edu/~treebank/home.html [Accessed 4 February 2014].

In a nutshell, then, Phase 2 of this study investigates written language, combines methods of frequency counts, concordances, keywords and collocation searches (see §5.9.1.5 below), and approaches the data in an exploratory way, together with two important aspects which were more qualitative: firstly, the categorisation of the different functions of each of the target verbs and, secondly, the identification of deviant uses of the selected MWUs and the errors giving rise to this deviance. Although there was no explicitly stated hypothesis, my informal hypothesis at the outset was that there would be differences between expert and student MWUs, and between use by students in the Literature and Law courses; this was based both on intuition and on the experience of working with these and similar students.

Finally, Phase 3 is quantitative in that it investigates the relationship between student writers' production of MWUs and their productive vocabulary size, as measured by the VLT, and their academic performance, as measured by their examination scores.

Phase 1 involved the collection of both numerical and biographical data from students who made up the sample − data which was essential to the aim of comparing different groups' vocabulary knowledge in terms of a number of variables − while Phase 2 is firmly grounded in the field of corpus linguistics. The last phase (Phase 3) brings together the data from the first two phases in various analyses.

5.4 METHODOLOGICAL RIGOUR

As Dörnyei (2007:48) observes, 'the basic definition of scientific research is that it is a "disciplined" inquiry, and therefore one thing research cannot afford is to be haphazard or lacking in rigour'. Such scientific rigour can be enhanced by ensuring that one's study conforms to the quality criteria of reliability and validity, although defining such 'rigour' is a highly contested and complex matter with 'little consensus' (Dörnyei2007:48–49). Dörnyei (2007:50) divides the discussion of these quality criteria into three parts: reliability, measurement validity and research validity. Put very simply, reliability is the extent to which the instruments and procedures that are used deliver the same results in the same population in different circumstances (Dörnyei 2007:50; Rasinger 2010:55). Measurement validity encompasses the concepts of the more traditional conceptualisation of validity, that is, criterion validity (the correlation of the particular test used with another, similar test), content validity (expert judgement about test content) and construct validity (the extent to which test results are consistent with the theory to which the 'target construct' belongs) (Dörnyei 2007:51).

Research validity goes beyond measurement validity in that it is concerned with the 'meaningfulness' of the interpretations placed on the findings of a study by the researcher (known as internal validity), and how far these interpretations can be generalised to different situations (external validity) (Dörnyei 2007:52).

Reliability and validity are important constructs in traditional quantitative and qualitative research, although they are assured in different ways according to the type of study being conducted. The sections which follow explain how these criteria were met in Phases 1 and 3 (§5.4.1) and in Phase 2 (§5.4.2).

5.4.1 Reliability and validity (Phases 1 and 3)

Phase 1 of this study made use of Laufer and Nation's (1995) Vocabulary Levels Test (VLT) (Active Version) to measure vocabulary size, and was quantitative in design. Laufer and Nation (1999:44-45) conclude that this test is a 'reliable, valid and practical measure of vocabulary growth' which 'allows researchers to investigate other aspects of vocabulary knowledge [than simply meaning and strength of vocabulary knowledge] and thus look more effectively at breadth of vocabulary knowledge'. However, as the test is one of *controlled productive ability* I have distinguished what it tests from what the corpus analysis reveals of students' word knowledge, referring to this as a test of vocabulary size, or breadth – that is, of how many words students know how to use in particular, controlled circumstances. The investigation of the depth of their word knowledge with regard to a specific set of high-frequency verbs used delexically in MWUs comes in Phase 2.

The VLT has been in the public domain for many years, and is freely available to researchers on the internet and in publications, with the result that its reliability and validity are generally accepted. Laufer and Nation (1999) report on a study they conducted to validate this test. They tested whether the VLT would distinguish 'among different levels of language proficiency since vocabulary size forms a part of language proficiency' (1999:38). Their results showed that the test was a valid measure of vocabulary growth because results reflected 'the gradual mastery of the successive frequency levels of the test as proficiency increases' (Laufer and Nation 1999:41). In terms of reliability, the entire test version A for all subjects had an internal consistency of .86 using the Kuder-Richardson formula KR21 (Laufer and Nation 1999:39).

These researchers also compiled another three parallel versions of the test, which they checked for equivalence by establishing whether they would correlate highly with one another when administered to the same group of learners (Laufer and Nation 1999:41). Correlations between the four versions were generally moderate to high and were significant. These tests also led to equivalent decision-making regarding the individual test takers. The authors found that any of the four versions of the test could be used for diagnostic tests, but for the purposes of test-retest situations they recommend two new parallel versions compiled from pairs of tests at each level which were not significantly different. Laufer and Nation (1999:44) report the reliability of parallel version C only, namely, 0.91 on KR21.

As noted above, Laufer and Nation (1999:44) found that this version of the VLT was a reliable, valid and practical measure of vocabulary growth. It can be printed on only three pages, which makes it economical

to administer, and it can also be computerised. In addition, it has a history of use in many research studies. As it is a modified cloze test, marking can be fairly objective. For each question, a 'meaningful sentence context' (Laufer and Nation 1999:37) is provided. The first few letters of the target word are supplied to prevent test takers from inserting a word which may be semantically appropriate but not from the same frequency level. Laufer and Nation (1999:37) used the minimum number of letters required to 'disambiguate the cue' (see Appendix A).

An independent marker marked the tests; where words could be comprehended easily, spelling mistakes were not marked as incorrect and grammatical mistakes were also ignored where the meaning was not obscured. Each learner was awarded six scores, one for each level of the test and the total for the whole test. The reliability coefficient (Cronbach's alpha), a test of the variance between these scores, in the main study was 0.934 (percentages)/0.816 (raw scores), which reflects a high level of consistency.

In Phase 3 of this study, after the VLT tests had been administered and the examination scores obtained, the link between breadth of vocabulary knowledge, as tested by the VLT, and the use of MWUs by students, was investigated by extracting a sample of texts from the student sub-corpora. Using examination scores as the organising variable, the ten students whose scores fell closest to the 25^{th}, 50^{th} and 75^{th} percentile groups respectively in both the Literature and Law groups were selected, resulting in a sample of 60 students in all. This is a form of non-probability sampling (Dörnyei 2007:98) known as quota sampling, in which the size of the sub-groups within a particular sampling frame is defined, and the sample is then selected from the group of subjects available; thus I had six sub-groups in total, with ten participants in each: these were students who achieved scores that were closest to the respective percentile. There was thus a limited group from which these particular participants could be chosen.

5.4.2 Rigour in corpus linguistics [Phase 2]

Various criteria are used to establish methodological rigour in the field of corpus linguistics. As Kennedy (1998:60) notes, 'issues in corpus design and compilation are fundamentally concerned with the validity and reliability of research based on a particular corpus, including whether that corpus can serve the purposes for which it was intended'. McEnery and Wilson (1996:22) sum it up as consisting of a sample that is 'maximally representative of the variety under examination' and 'is finite in size'. It is also in electronic format so that it can be read by a computer and should represent a 'standard reference for the language variety which it represents' (McEnery and Wilson 1996:22). In the following subsections, important methodological concepts which help to ensure validity and reliability in corpus linguistics are discussed in more detail.

5.4.2.1 Representivity and balance

Gilquin and Gries (2009:6) observe that what is important to the validity of a corpus study is that the corpus is representative and balanced, containing data for each part of the variety/register/genre it is supposed to represent. Tognini-Bonelli (2001:57) sums it up thus:

> There seems to be general agreement among scholars who choose to work on a corpus that this should be representative of a certain population and that the statements derived from the analysis of the corpus will be largely applicable to a larger sample of the language as a whole.

In other words, the results obtained from the analysis of the Student corpus in this study should be applicable to students of English at any South African university; the texts making up this corpus can be regarded as representative of the courses and population under study, comprising as they do writing by students studying English at a South African university through the medium of English. These students come from a range of language backgrounds including native, non-native and foreign language speakers of English. Although this university and others like it in South Africa use English as the medium of instruction in the main, this is not the language spoken by the majority of students in their day-to-day conversation and private lives.

The Expert corpus, on the other hand, comprises writing by expert writers of English – also both native and non-native speakers – writing that is typical of study material at any English-medium university in South Africa. The writing in this corpus is regarded as a model of the sort of academic writing that undergraduates at this level would wish to emulate.

Gilquin and Gries (2009) also believe that it is important that a corpus is balanced in that the parts of the corpus are proportional in size to the parts of the variety/register/genre the corpus represents, and that it contains texts that have been produced in a 'natural communicative setting' (2009:6). The corpora in the present study can be regarded as fairly typical in that they are representative of the genre (student academic writing in two specific areas of English and expert academic writing from the same two fields), and the Student corpus comprises language produced in a natural communicative setting, although this writing was produced under timed conditions. Gilquin and Gries (2009:7) observe in this regard that writing examination essays under time constraints is natural in the university classroom context. The corpora are balanced in that they broadly represent one register – academic writing (although student academic writing and lecturer academic writing are somewhat different in systematic ways) – and both are split into two sub-corpora according to the courses of English Literature and English Communication for Law. These courses were chosen firstly because they represented two different academic genres within the discipline of English studies and I wished to investigate course- or genre-related differences in student writing. Secondly, as I

work in the department where both these courses were housed I had easy access to both students and course material. Thirdly, I was teaching or had taught all three modules and I was thus familiar with the content and the approach to teaching both courses. The choice of these two variants of academic English is well supported by increasing recognition that the disciplinary variations in academic genres can provide researchers with useful insights (Lee and Chen 2009:150), and Hyland (2008) and Hyland and Tse (2007) have suggested that focusing on specific texts from a specific genre can be more rewarding pedagogically than analysing general academic English.

5.4.2.2 Situational and linguistic criteria

Tognini-Bonelli (2001:60–61) notes that situational criteria (register and genre) and linguistic criteria (the selected linguistic features of the texts making up the corpus) are not independent of each other, and that one of the main uses of a corpus is to investigate relationships between them. She follows Biber (1994, cited in Tognini-Bonelli 2001:61) in suggesting that the best route to take is to base the initial selection of texts on situational parameters – these can be identified in advance while linguistic types cannot. In the present study, I was confined to the writing of those students who had completed the VLT, although I based the initial selection of students on the courses in which they were enrolled for the reasons outlined above (see §5.4.2.1). Register was also predetermined by the fact that the corpora comprised academic writing. The same criteria applied to the compilation of the Expert corpus in that I used all the available writing pertaining to the two courses in question. In this way the situational and linguistic criteria were as rigorously adhered to as was possible in the circumstances, strengthening the validity of the study.

5.4.2.3 Size of the corpus

One of the most controversial issues related to the validity of corpus analysis is the size of the sample of language in a corpus. This issue has more to do with 'the theoretical stand on the nature of language than with computer storage' (Tognini-Bonelli 2001:62). Stubbs (1993:11) believes that the unit of study must be whole texts; 'few linguistic features of a text are distributed evenly throughout'. Sinclair (1991:19) supports this – 'a corpus made up of whole documents is open to a wider range of linguistic studies than a collection of short samples'.

The corpora in this study were situationally defined and, as such, I had to work with the data I had. As the Student corpus is a learner corpus the texts are necessarily short, albeit 'whole' and authentic, because they were written by students under examination conditions with definite time constraints. Despite the fact that texts ranged in length from as few as 130 words to as many as over 2000, I anticipated that this corpus would reveal interesting as well as significant information about students' use of MWUs. The Expert corpus, on the other hand, consists of texts which vary in length from a few hundred words to several thousand.

As to the adequacy of these corpora in providing valid and reliable results, as Kennedy (1998:68) puts it, 'a corpus is more or less adequate according to the extent to which [it] matches the purposes to which it is put'. Exactly how big a corpus needs to be in order to do this is still the subject of some debate; various researchers have made claims about how large corpora should be to provide significant results. Kjellmer (1991:115), for instance, suggests that a corpus should be at least a million words in size to 'contain a considerable part of the English phrases in current use', while Barkema (1993, cited in Kennedy 1998:117) in his study revealed that even 20 million words was not enough to provide sufficient examples for research into particular collocations.

Although some of the best known and best documented studies in the area of computer-assisted corpus analysis, or concordancing, in language teaching and learning have been done using huge corpora, for example by Sinclair and his colleagues on the COBUILD Project (Sinclair 1991, §2.2.2.2), many researchers have worked with small, self-compiled corpora of L2 or learner writing which have yielded very interesting findings. These include Granger (1998b), Laufer and Waldman (2011) (a learner corpus of about 300 000 words), Lee and Chen (2009) (CAWE: 407 960 words; BAWE-L: 177 153 words; EXJA: 388 490 words), Nesselhauf (2005) (data from sub-corpus of ICLE [German]: 32 essays, average length 500 words), Pienaar and De Klerk (2009) (ISAE corpus: 60 000 words), Van Rooy (2006:46) (three comparable corpora from ICLE representing 'inner', 'outer' and 'expanding' varieties and the TLE corpus and the LOCNESS for comparison, all of which were about 200 000 words in length) and Van Rooy and Terblanche (2006) (two corpora of just over 200 000 words each: TLEC: 519 essays, average length 400 words; LOCNESS: 166 essays, average length 1200 words). These research studies would seem to justify the use of the relatively small corpora in the present study; as far as size of the corpora used in studies is concerned, then, there seems to be no hard and fast rule. As Pienaar and De Klerk (2009:358) observe, '[t]here is no ideal corpus size, only an optimum corpus size determined by the research needs and pragmatic considerations such as the availability of resources'.

This section has provided an explanation of the steps taken in this study to ensure the discipline of the enquiry and the 'legitimacy' of the findings (Dörnyei 2007:48). The following sections discuss the procedures followed in the study, starting with the pilot study.

5.5 PILOT STUDY

Dörnyei (2007:75) likens a pilot study to a dress rehearsal and stresses that it is essential to pilot one's instruments and procedures in order to ensure the reliability and validity of the results of one's project. Because quantitative studies 'rely on the psychometric properties of the research instruments', such as assessments or questionnaires, piloting is essential to ensure that all variables have been adequately

covered by the questions (Dörnyei 2007:75). Qualitative studies do not require that same sort of piloting, but techniques to analyse the data, for instance, can be trialled before the main study begins. A pilot study can iron out problems at the outset and obviate later difficulties.

In this case, the pilot study was conducted in 2008. The main purpose of this pilot was to test the viability of the data collection method and the instrument to be used in Phase 1. Collecting the data was a complex and potentially expensive issue and in order to secure funding from the Dean's research fund for the process it was essential to test it out initially on a small scale.

Participants

I targeted students enrolled in two first-year English Literature modules and an English for Law module who were attending lectures at the main campus in Pretoria. This was one session of lectures that were offered once a semester at the main centres in South Africa (Pretoria, Polokwane, Durban and Cape Town) and attendance was voluntary. This group can be regarded as constituting a representative sample of students enrolled in these courses. I used convenience sampling as these students were available together in one place and willing to complete the test. The sample comprised 132 ENN106J (*English Communication for Law*) students, 37 ENN101D (*English Studies: Approaching Literature and Writing*) and 85 ENN102E (*English Studies: Explorations in Reading and Meaning*) students (a total of 122 literature students in all, as these two groups were later combined) (see §5.4.2.1).

Vocabulary test and questionnaires

The test I used was Version 2 of Nation's (1983, 1990) Vocabulary Levels Test (Schmitt et al. 2001) as it tests five levels of vocabulary, namely the 2000-, 3000-, 5000-, and 10 000-word levels and academic vocabulary and has been used as a predictor of academic success (measured in this study by examination results). This version of the VLT tests receptive (or passive) word knowledge by requiring students to choose a word from a list of possibilities in order to complete a given sentence. In other words, it tests whether students recognise the word and understand its meaning well enough to select it for use in a particular context. It does not, however, test whether students can produce the word in the correct context.

At the same time, I asked students to complete a questionnaire eliciting biographical data such as information on the course in which they were enrolled. Further information on their gender, age, language background and ethnicity was obtained from their registration records.

Procedures

The test was administered in September 2008. In accordance with ethical procedures, students attending the lectures were given a letter explaining the nature of the study and providing evidence that the

researcher had permission from the university to conduct it (see Appendix B). They were also given the option not to take part. Those students who were willing (almost all the students present) completed the test and the questionnaire. Students took between 10 and 35 minutes to complete the test.

Coding

I marked the tests myself and captured the data using SPSS (IBM SPSS Statistics version 19), a program which allows the average researcher to engage in statistical analysis quite easily, using a personal computer. Once the end-of-year examination results had been released, these students' percentages were also captured. As 51 of the original group did not write the exam, the final sample comprised 203 students.

Results

Table 5.1 reflects the students' results on the VLT and the examination. As can be seen, the general trend in these results is that the students were performing very well at all levels, except at the 10 000 word level. Even students at the 25th percentile of the VLT results had achieved mastery of the first two levels (i.e. 85% and above) and were close to this at Level 4. This suggested that the test had not discriminated well between students and had produced a fairly 'flat' profile – either this or students' receptive vocabulary knowledge was on average very good, which was certainly not reflected in their examination results. These results could also be regarded as being rather skewed, since more motivated students are more inclined to attend extra lectures when these are offered.

	Level 1	**Level 2**	**Level 3**	**Level 4 UWL**	**Level 5**	**Total**	**Exam**
	2000	**3000**	**5000**		**10 000**		
N = 203							
Mean	92.71	90.44	78.05	81.76	43.09	76.94	52.53
Sd	13.968	15.742	20.286	23.327	29.021	16.530	12.68
							19 –
25th percentile	90.00	88.89	66.67	73.33	23.33	70.07	45.00
50th percentile	96.67	96.30	83.33	90.00	36.67	78.23	53.00
75th percentile	100.00	100.00	96.67	96.67	60.00	87.76	62.00

Table 5.1: Pilot study: Mean scores on Nation's Receptive Vocabulary Levels Test (VLT) (1983, 1990) and examination

The mean scores on the whole vocabulary test and the exam were then calculated according to students' course of study. These scores are reflected in Table 5.2:

	Lit Vocab	Law Vocab		Lit Exam	Law Exam
	N = 92	N = 111		N = 92	N = 111
Vocab mean	82.74	72.13	**Exam mean**	51.33	53.53
sd	14.198	16.837	**sd**	12.448	12.845
					19 – 77
25th percentile	74.49	67.35	**25th percentile**	45.00	50.00
50th percentile	82.99	74.15	**50th percentile**	50.50	54.00
75th percentile	95.07	82.99	**75th percentile**	59.00	63.00

Table 5.2: Pilot study: Mean scores on Nation's Receptive Vocabulary Levels Test (VLT) (1983, 1990) and examination per course

As is clear from Table 5.2 above, the Literature students performed better than the Law students on the VLT. Using the Pearson correlation technique, there was a modest positive correlation between students' examination scores and academic vocabulary scores ($r = 0.325$, $p < 0.0005$).

The lack of discrimination in the test of receptive vocabulary, as reflected in both Tables 5.1 and 5.2 above, led to the decision to replace it with an 'active version' of the VLT (devised by Laufer and Nation in 1995 and revised in 1997) in the main study (see §5.6). As my assessment of vocabulary depth was based on written production, it was also more appropriate to use the productive version of the test for my vocabulary breadth measures. This latter version of the VLT tests the same five levels of vocabulary (2000-, 3000-, 5000-word levels, academic vocabulary and the 10 000-word level) but requires students to complete words rather than simply to identify them; that is, it is a modified cloze test. I conducted this test among 16 Literature students at the Cape Town lectures in April the following semester. The results of this test are reflected in Table 5.3 below.

	N	2000-word	3000-word	5000-word	UWL	10 000-word	Total %	Exam %
Mean	*16*	97.43	84.72	71.88	74.65	50.69	75.72	55.67
Std. Dev		4.79	16.79	30.62	21.37	34.06	20.35	13.94
Min-max		88–100	50–100	6–100	38–100	0–100	39–97.70	28–75
Percentiles								
25th		95.58	79.16	46.87	52.77	11.11	56.89	45
50th		100	88.88	84.37		58.33	83.33	57
75th		100	98.61	98.43	94.44	77.77	93.38	68

Table 5.3: Pilot study using Active version of Vocabulary Levels Test: Mean scores on levels test and examination

Given the small numbers in each group, a non-parametric correlation, Spearman's rho, was used. Spearman rho correlations revealed a very significant strong correlation between scores at the UWL level and the total vocabulary score (r_s = .85, p < 0.01) and between UWL scores and exam scores (r_s = .88, p < 0.01). This suggested that there was a strong relationship between students' productive vocabulary size, both in general and in terms of knowledge of academic words, and academic performance. Following these results, I went ahead with the data collection for the main study in 2010, using the active version of the VLT.

5.6 MAIN STUDY

The main study comprised three phases (see §5.2) with Phase 1 being conducted during the first semester of 2010. The methodological aspects of the study that apply to all three phases are discussed first; the three phases are then discussed one by one, in separate sections.

5.6.1 Participants

All 4500 students enrolled for these modules were sent the active version of the VLT and the questionnaire by post at the beginning of the first semester in 2010. The 346 students (161 enrolled in the Literature modules and 185 in the Law module) who made up the initial sample were identified by the fact that they completed and returned the test and questionnaire (an 8% return rate). Of these 346 students, a total of 298 wrote the examination at the end of the semester. The scripts of these 298 students were identified and their examination essays were transcribed electronically and saved as Word documents. As was the case in the pilot study, the method used was convenience sampling: the sample was made up of those students who chose to return the questionnaires and tests and who then sat for the final examination in their respective modules. Although a common challenge in postal questionnaire administration, this last point suggests a limitation to this study: the sample was in effect self-selected, implying in this case that it was probably those students who were on average above the norm in their group in terms of commitment to their studies and motivation, possibly also in terms of performance, who were more likely to complete and return the test. This is an issue that needs to be borne in mind when assessing the internal validity of the study, and also the generalisability of its findings, which relates to external validity. In addition, the test was completed at home without any invigilation, which may have skewed results. However, the fact that many students performed poorly in the vocabulary test in the main study, despite the fact that they were not supervised when writing it suggests that the results may actually reflect students' vocabulary knowledge fairly realistically. Scores were also all consistently lower than the pilot study receptive results, which is probably to be expected because of greater difficulty in the active test. This is discussed in more detail in Chapter 7 (see §7.5).

Of the initial sample of 346 students, almost three-fifths (57.9%) indicated in their questionnaires that their home language was not English. Of this 57.9%, 41.1% were mother tongue speakers of an African language. A quarter (25.5%) of those who indicated that their home language was English were, according to their university registration information, black students. This may be evidence of what Coetzee-Van Rooy (2012:88) refers to as 'widening language repertoires where English is used in addition to other languages'. In her study, Coetzee-Van Rooy (2012) investigated the language repertoires of students at a South African university, in an attempt to cast more light on the varying views on the perceived shift of speakers of African languages towards English. Her findings revealed extensive multilingualism amongst her participants. Perhaps the most pertinent finding of her study to mine is her finding that many students reported their home language as an African language, but their strongest language as English. Although she notes that this could be seen as evidence of a potential language shift to English among these participants, all her other findings seem to suggest that it is evidence rather of a 'flourishing functional multilingualism' (Coetzee-Van Rooy 2012:112). That is, for the students in her study, the home language functioned in 'the domain of the family' (Coetzee-Van Rooy 2012:112), while English was used for reading and writing.

In a later study, Coetzee-Van Rooy (2014) interviewed some of the students who had taken part in the language repertoire study (Coetzee-Van Rooy 2012) in order to try to further explain this 'ordinary magic of stable African multilingualism' (2014:121). She found that those interviewed generally viewed multilingualism as very probable, confirming Coetzee-Van Rooy's (2014:126) impression that these students functioned in a 'multilingual language mode'. In several instances, participants equated their home language with their sense of their own identity, but also provided support for the notion that in order to fit into South African urban society one has to be multilingual; they expressed an awareness of the importance of English, while at the same time maintaining their home language (2014:129). The interviews confirmed the finding from the questionnaire data of a 'functional distinction' (Coetzee-Van Rooy 2014:133) between those languages that are used socially and those used at school or university, although some participants did express an awareness of the strangeness of the fact that they were not equally proficient in reading and writing in both English and their home language.

These studies by Coetzee-van Rooy (2012, 2014) have revealed something of the 'complexity' of the phenomenon of multilingualism in South Africa (Coetzee-van Rooy 2012:113). This issue has already been discussed in some detail in Chapter 1 (see §1.2).

Students in this sample were expected to fall into the 'intermediate' or 'proficient learner' category (according to the research done in the National Benchmark Test Project[21] [2009]), having had at least nine

[21] 'Proficient: Performance in domain areas suggests that academic performance will not be adversely affected. If admitted, students should be placed on regular programmes of study. Intermediate: Challenges in domain areas identified such that it is

years, from Grade 4 to 12, with English as LoLT (language of learning and teaching). Students ranged from first-year students to those who had registered for the first time as many as thirty years before. Their ages ranged from 18 to 68, with 20% under the age of 20 at the time of the study. Almost half of the sample (48.5%) was under 40 years of age. It is thus clear that the student body from which the sample was drawn was more mature and more diverse in age than would probably be the case at a residential university. Registration information listed the ethnic group of the majority (57.0%) as 'black', 30.9% were 'white' and the remainder 'coloured' or 'Asian'. The reason for including this information was to form a better idea of the varieties of English spoken by these students. Women outnumbered men, with the former making up almost three-fifths of the sample (59.3%).

The student population at this university is multicultural and multilingual, but the medium of instruction is English and, in order to pass their degree, students must be able to read and write in English at a fairly advanced level. This sample, though small relative to the enrolment numbers at the university (approaching 400 000), can be reliably regarded as representative of students registered in the Human Sciences at any South African university. It includes 'the full range of variability in a population' (Biber 1994, cited in Tognini-Bonelli 2001:59), representing speakers of all 11 languages which have official status in South Africa, and foreign languages besides. The sample is taken from a specific population (students in the College of Human Sciences) but the sample can be said to include the full range of variability within that small target population. As this is an open and distance e-learning institution, many students are adults who have come late to higher education, but there are many, too, who are straight out of high school with very little life experience.

As noted above, in the end only 298 of these students wrote the examination; for some reason or another, 48 (13.8%) of the original group did not sit for their final examination. The final sample in this study was thus made up of 298 students, 139 enrolled in the Literature modules and 159 in the Law course.

5.7 PHASE 1: VOCABULARY SIZE AND ACADEMIC PERFORMANCE

Phase 1 was driven by the first two main research questions (see §5.2), measuring the size of students' productive vocabulary in relation to course, gender, age and language background, and exploring the relationship between vocabulary size and academic performance.

predicted that academic progress will be affected. If admitted, students' educational needs should be met in a way deemed appropriate by the institution (e.g. extended or augmented programmes, special skills provision) (National Benchmark Test Project 2009).

5.7.1 Research instruments

In order to collect data for quantitative analysis, a questionnaire and a vocabulary test instrument were used in this phase. Students' examination results were used as a measure of academic performance.

Questionnaire

The questionnaire was designed to elicit biographical data only (see Appendix A) – student number and course code and language background (that is, home language). Students were also asked to indicate when they had started and ended the test. This detail was included to elicit information for purposes of comparison of students' scores and rate of completion, and also to ensure that students had completed the test at one sitting. In the event, this information was discarded as too few students completed this part of the questionnaire.

Further information – gender, age and ethnicity – was taken from the information provided by each student on registration and accessed via individuals' student numbers. The university is required by the state to provide statistics on ethnicity and, for the purpose of this study, information on this and on language background was included to allow for the identification of any particular idiosyncrasies in student writing which might be the result of L1 transfer or of the variety of English spoken. The inclusion of language background also meant that NSs and NNSs of English could be identified in the sample, as Unisa has a heterogeneous student body with speakers of all 11 official languages as well as many foreign languages. This study compares student writers with expert writers and, although the L1 speakers of English in the student sample were not 'learners' in the sense of 'language learners', they were novice writers of English in the academic idiom. It should be noted that this study is not primarily concerned with the influence of students' specific language backgrounds, although this could be the focus of a further study.

Laufer and Nation's (1995) Vocabulary Levels Test (Active Version)

Students' productive vocabulary was measured by means of a word completion exercise, a modified cloze test, an active version of the VLT originally devised by Nation (1983) and used by Laufer and Nation (1995) (see Appendix A). This VLT is freely available to researchers and both the passive and active versions have been used or adapted in many studies (among others Laufer and Paribakht 1998; Laufer and Nation 1999; Lenko-Szymanska 2000). This instrument tests five levels of vocabulary, that is, the 2000-, 3000- and 5000-word levels, the UWL and the 10 000-word level. Such word-completion exercises as this test contains are associated with the ability to use vocabulary productively (Laufer and Nation 1999). I accordingly used the test in this study to create a profile of students' productive word knowledge.

End-of-semester examination marks

In order to establish whether there was a relationship between vocabulary size and academic performance, that is, whether high scores on the VLT would be linked to higher examination scores, and vice versa, students' marks in the end-of-semester examination in the English and Law courses they were enrolled in were collected. This was the same semester as that in which they had completed the VLT.

5.7.2 Procedures

At the beginning of the first semester of 2010 (which ran from January to June), having received permission from the university to conduct the study (see Appendix B), I sent out the questionnaire and the test, together with a covering letter explaining the project, to all 4500 students registered for the three modules, English Literature studies (ENN101D and ENN102E) and English Communication for Law Students (ENN106J). The covering letter explained the project to students and requested that they fill in the questionnaire and complete the vocabulary test, without using a dictionary (see Appendix A). They were asked to send both questionnaire and vocabulary test back to the university in the stamped addressed envelope provided once they had completed them. Those students who returned these documents also gave their permission for their results and examination scripts to be used in the study. Completed questionnaires were returned to the university by 346 students.

5.7.3 Scoring and analysis

In this first phase of the study, students' responses to the questionnaire and vocabulary test were marked, scored and analysed. The information gathered from the questionnaires was saved in SPSS and all computations were made using this program. The biographical data were used to create a profile of the sample according to course of study, gender, age and language background.

Once captured in SPSS, the test scores were computed using descriptive and inferential statistics. ANOVAs (to test for significant differences) and correlations (to test for relationships) between vocabulary results and students' examination scores were performed. Two correlation measures were used: the Pearson product-moment parametric test of correlation, a measure of association between two continuous variables, where the data is normally distributed, and Spearman's rank order correlation rho, a nonparametric test of correlation, used when the group sizes were small and assumptions about normal distribution could thus not be met.

Multiple regression analysis was also performed to identify which levels of the VLT would best predict academic performance, using the Enter[22] method (Brace et al. 2003:214). Multiple regression is an

[22] The 'Enter' method enters all variables into the equation at the beginning. The 'Stepwise' method only enters the variables which are good predictors into the final equation.'

extension of simple linear regression and is a statistical technique that allows a researcher to predict a subject's score on one variable on the basis of his or her scores on several other variables. 'The term "predictor variables" refers to those variables that may be useful in predicting the scores on another variable, referred to as the "criterion variable"' (Brace et al. 2003:210). In other words, 'multiple regression allows us to identify a set of predictor variables which together provide a useful estimate of a participant's likely score on a criterion variable' (Brace et al. 2003:214).

5.8 PHASE 2: COMPARISON OF STUDENT AND EXPERT WRITERS' USE OF SELECTED VERBS AND MWUS WITHIN AND ACROSS ACADEMIC GENRES

Phase 2 was driven by two research questions: the first related to the distribution of the functions of the three selected verbs in the Expert and Student corpora, and the second to the use of selected MWUs in the two corpora (see §5.2).This second phase of the study comprised the delimitation of the MWUs under study, and the compilation and analysis of the corpora. In this section, the compilation of the corpora is discussed, followed by the formulation of the analytical framework used to analyse the MWUs.

The collection of the corpus data was informed by the theoretical thinking current in corpus linguistics and in the area of research dealing with MWUs. The analysis of the data subsequently revealed the relevant linguistic characteristics of the writing making up the Student corpus, and how these compared with the Expert corpus.

In an attempt to fulfil the methodological requirements suggested by Tognini-Bonelli (2001), Nation (2001) and Gilquin and Gries (2009), this study has set out clearly the research questions (see §5.2), collected two corpora, a reference corpus of expert writing and a primary research corpus of student writing (see §5.8.1 below), which are in electronic format and machine readable, representative of the genre under investigation, and collected in a natural, communicative setting. A reputable and proven computer software program, WordSmith Tools Version 5 (Scott 2008) was used to organise and analyse the data (see §5.8.1.5).

5.8.1 Compiling the corpora

The corpora investigated in this study are what Baker (2010) would call 'specialised' in that they contain a specific type of language, namely, essays written by students under examination conditions on the one hand, and study material written by expert lecturers on the other, consequently involving inevitable restrictions on size and genre. As this study compares the use of combinations of words in English, no other

language is included. The Student corpus consists of student academic writing in English, much of it by NNS students. This is a diverse group, however, with wide variations in terms of age, language and educational background. Such a corpus should reveal correct usage, as well as common errors, areas of difficulty, and 'over- and underuse of lexis or grammar when compared to an equivalent corpus of native speaker language' (Baker 2010:100), in this case the Expert corpus. This latter corpus comprised expert writing obtained from the study material used in these Literature and Law courses (see Appendix C).

5.8.1.1 The Student corpus

The Student corpus was split into two sub-corpora, Student Lit, which contained examination writing from students enrolled in the two Literature modules, and Student Law, containing examination writing by students enrolled in the English Communication for Law course.

Although I took cognisance of points made by various experts in this field in the corpus compilation process (see §5.4.2), circumstances dictated that I deviate on several fronts from the recommendations of researchers such as Biber and Tognini-Bonelli. As a teacher at a distance education institution, the only opportunity I had to elicit authentic writing from the study population was during examination time. This is the only time that students are seated in a venue writing their own, authentic responses to a question or questions – albeit in different venues all over the world. Examinations are also the only time that students write under controlled conditions, with no access to reference material, the internet or other possible input such as tutor instruction. All other work they submit during the semester is completed beyond the confines of the university or its satellites and as such cannot be guaranteed as authentic. Because of this, I was bound to use the essays written in students' examination scripts, some of which were rather brief; these can, however, be regarded as 'complete texts'. This, I believe, makes these corpora authentic.

In this study, the Student corpus is made up of a total of 206 173 words from 298 essays (139 Literature and 159 Law) written under examination conditions by a group of undergraduates. The rationale behind the compilation of this corpus was that the study focuses on a very specific group of participants (the sample drawn from the wider target population, see §5.6.2). When split into corpora according to genre, the Student Lit corpus comprised 142 655 words and the Student Law, 63 518. As there were no length stipulations for texts other than the requirements of the examination, that is, between 350 and 450 words per essay, and the time constraints of the examination, two hours in this case, texts varied quite dramatically in length among students and across courses, from 350 to over 2000 words in the Student Lit corpus and from 131 to over 800 words in the Student Law corpus. Literature students wrote an average of 1026 words in their exam scripts, while Law students averaged 399 words. However, as the focus of the study is on the group of students rather than the individual, this was deemed acceptable.

The kinds of texts selected were controlled by the examination requirements in the modules concerned – academic argumentative or discussion-type essays on either a legal type of question in the Law module (*Immigration has contributed to a rise in international crime* or *Corruption is mainly practised by the rich* or *Laws are punitive and not corrective*) or a question on a prescribed novel, poem or extract from a reader in the Literature course (see Appendix D for examples of the examination papers). The two sub-corpora are not identical in size – this was an unavoidable and unforeseen result of the fact that the students in the Law course wrote much shorter essays in their examinations than had been anticipated.

The Literature examination papers required students to answer two questions in total. There were five questions in the paper for ENN101D: poetry (comprehension-type questions set on a poem); comprehension questions on a passage from a novel; and an essay question on each of three novels, *Nervous Conditions*, *Heart of Darkness* and *The Madonna of Excelsior*. Even though there were two questions featuring comprehension-type questions, students were nonetheless required to write extended texts in answer to these questions. The question paper for ENN102E followed the same format, except that there were only four questions to choose from, there being no comprehension-type question: poetry, *Disgrace*, *The Great Gatsby* and *The Merchant of Venice* were examined. I used all answers in the scripts, whether they were answers to essay or comprehension questions.

The question paper for the Law students comprised multiple-choice questions set on a legal case, and an argumentative essay. Only the essays were used for the corpus. There were three essay questions to choose from and students were required to write at least 450 words. A possible limitation here is that more marks, and potentially more time, were allocated to essays in the Literature examination papers; in the Law exam, the essay counted only half the marks towards the total, and the first section, the multiple-choice test, possibly took more than half of the allocated time to answer. This could also be an explanation for the brevity of the texts.

5.8.1.2 The Expert corpus

The Expert corpus, used as the reference corpus, was compiled from the study material and tutorial letters sent to students enrolled for the Literature and Law courses during the semester of study. This was made up of 25 texts in all, 21 Literature texts and four Law texts, with a total of 192 060 words (see Appendix C for a list of texts). This material included study guides for all three modules, tutorial letters containing general information, specific information and instructional material on prescribed works and course content, as well as supplementary teaching material in the form of feedback letters on assignments, extra notes on prescribed works and so on. The fact that there were so many more Literature than Law texts reflects differences in the sort of teaching conducted in the two courses; students in the Literature courses were required to read more and had more options to choose from in their assignments. As a result, they received more feedback letters. Students in these courses were also given the opportunity to attend

lectures once a semester in the main centres in the country. Notes from these lectures were then sent to all students registered for the course to accommodate those who had been unable to attend the lectures. The Law students were not given this opportunity.

This writing, that is, this feedback to students from lecturers, is regarded as authentic and representative of expert academic writing, having been composed by several academics, experts in the fields of English Literature or English Communication for Law, and having undergone stringent editing, peer review and critical reading. The writing making up this corpus was written between 2004 and 2010 and was thus considered appropriate. I was part of the team responsible for some of these texts but as all of them were written before I started this study this had no influence on the outcomes. For purposes of comparison, this corpus was split into two according to course, resulting in Expert Lit (144 231) and Expert Law (47 829 words) sub-corpora.

Table 5.4 reflects the composition of the corpora:

	No of Texts	**Tokens**	**Mean Tokens per text**	**Types**
Expert Corpus	25	192060	7682	11595
Student Corpus	298	206173	691	9682
Expert Lit	21	144231	6868	10094
Expert Law	4	47829	11957	4867
Student Lit	139	142655	1026	7475
Student Law	159	63518	399	5255

Table 5.4: Composition of the corpora

5.8.1.3 Characteristics of the corpora

As far as the situational criteria of register and genre are concerned (see §5.4.2.2 above), the texts making up the corpora in this study can be regarded as representative of an academic register and as belonging to the genre of written academic language. As such, the text type (that is, the linguistic perspective) should be formal and impersonal in the main rather than colloquial and interactive. In other words, as the Student corpus was derived from writing of an academic and formal genre produced under examination conditions, it was expected not to feature many elements of spoken language, such as contractions, colloquialisms or personal pronouns (Pennebaker 2011). As the study guides are designed to be interactive, in the sense that they engage students in activities and attempt to fulfil the role a lecturer would play in a conventional university lecture, the Expert corpus could, as a result, be expected to contain, in addition to formal academic writing, more instructional language and the use of more personal pronouns.

As to being representative, texts making up the Student corpus had to be taken from the sample's examination scripts, as explained above. I could not control for the length of students' texts and had to make do with whatever they wrote (see §5.4.2.3).

5.8.1.4 Preparation of the corpora

All the examination scripts were photocopied and typed. The electronic copies were checked against the photocopies of the originals to ensure that the typing was absolutely accurate. In the process, each essay was coded with the student number and course code, as indicated on the questionnaire by students, and a code for the essay question or questions answered by the student. An example of the coding is <12345678_ENN101D_P_NC>, where *12345678* is the student number, *ENN101D* the course code and *P* and *NC* indicate the type of question, such as *P* for *Poetry* and *NC* for *Nervous Conditions*. These are two of the five prescribed texts in this module on which students were examined (see §5.8.1.1 above). The angle brackets were used to separate this coding from the text, as WordSmith Tools (WST) can be set to ignore the contents of these brackets. This means that these codes did not show up in the results of the analysis but could be used by the researcher to identify certain groups or individual texts within the corpus. Although students' numbers were never linked to names of individuals in this study, it was necessary to include these numbers in order to calculate the correlations between students' lexical proficiency, as indicated by their vocabulary size, their examination scores and their use of delexical MWUs in Phase 3. As no comparable analysis was needed for the Expert corpus, no coding was applied to it (and WST automatically keeps texts separate).

5.8.1.5 Analysis of corpus data

Five steps were followed in the analysis of the corpora, using WST (version 5). These are discussed in detail below. WST is a suite of computer programs (Scott 2008) which Scott describes as a 'Swiss army knife' (2001:48). This is an appropriate analogy as it captures Scott's intentions in designing the program – it was devised with ease and convenience in mind for the ordinary researcher, language teacher or student to use on a standard personal computer. Scott aimed to 'produce tools which would be general purpose in nature as opposed to specific' (2001:48). In order to comply with these requirements it was important that the program could be run using standard equipment and standard texts and that an affordable amount of computer memory and Windows would be adequate to run it. He designed it 'to be able to handle virtually any text or corpus in more or less any language' (Scott 2001:49). Language in the corpora would not need to be annotated or 'marked up', although it was designed also to deal with tagged text, and to be able to separate marked-up text from the text proper.

Following Wang and Shaw (2008), the five steps in the process were as follows:

Step 1 – generation of wordlists

Wordlists were generated for both corpora – in their entirety and also as sub-corpora – using the WordList application of WST. This allows the researcher to create lists of words ordered by frequency and alphabetically, both for the whole corpus and for the individual texts it comprises. In the present study, although individual word frequency was of less interest than multiword items, it was important to determine whether the identified verbs in question –*HAVE*, *MAKE* and *TAKE* – were in fact frequent in these corpora and this was therefore established at the outset. WordList was also used to establish the number of types and tokens in each corpus.

Step 2 – generation of keywords

Keyword lists revealed which words were significantly more frequent (in my study, significance was set at $p \leq 0.05$) and those which were significantly less frequent in the Student corpus compared to the Expert corpus as a whole, and in the Student sub-corpora when compared to the relevant Expert sub-corpora. There must be two lists available for the Keyword (KW) application to be able to generate a list of 'keywords'. The second step was thus to compare the entire Student corpus with the Expert corpus, and then the Student Lit and Law sub-corpora with the Expert Lit and Law sub-corpora respectively by generating keyword lists.

The reference corpus should be 'an appropriate sample' of the language in which the text under study (the 'node' text) is written (Scott and Tribble 2006:58). In this case 'appropriate' usually means a corpus which is fairly large, probably containing many thousands, even millions, of words (Scott and Tribble 2006). In the present study the Expert corpus was used as the reference corpus, because this was the benchmark against which the Student corpus was investigated. This was considered acceptable by an expert in the field of statistics and in working with WST.[23]

Step 3 – generation of concordance lines

The Concord application of WST was used in the generation of concordance lines. This tool allows the researcher to find all references to a given word or phrase within the corpus, and it shows these in standard concordance lines with the search word centred and a variable number of words – co-text, commonly referred to by researchers as context – on either side of it. Searches can be refined by searching again on the first word to the right, or the second to the left and so on, allowing the researcher to form a very good idea of how the word behaves in the corpus and 'the company it keeps' (Firth 1957, cited in Léon 2007:1). A

[23] Personal correspondence with Prof. D. Prinsloo, University of Pretoria.

separate concordance was generated for each word-form of each selected verb e.g., *have, has, had* and so on (what Biber [2006:34] refers to as the 'inflectional morphemes' of the base word or lemma).

Thus Concord provides information on a word's collocation as well as its colligation, or its tendency to have a 'grammatical prosody' (Scott 2001:54), or a typical grammatical environment, by showing up patterns and clusters featuring a particular word or phrase, such as

his times. In other words, he **has** doubts about the stereotype (Exp Lit)

Step 4 – investigation of concordance lines

In order to identify all occurrences of these three verbs, concordance lines were generated for each word-form of each verb, e.g. *hav*, has, had*. The forms *hav**, *mak** and *tak**, called 'wild cards' in corpus linguistics, were used instead of *have*, *make* and *take*, as the use of the asterisk identifies all words starting with the three letters and allows the researcher to identify forms such as *having*, *haven't*, *making*, *taking* and *taken*. The use of such wild cards can, however, generate anomalies and did so in this study. These anomalies were excluded by 'cleaning' the data before statistical analyses were performed. This step was completed by using the Concord application of WordSmith Tools (§5.9.1.5). In this respect, the Concord function was arguably the element of the program most useful to this study in that it allowed me to focus on how a word or phrase was used or behaved in context, or as Flowerdew (1998:542) observes, 'in naturally occurring text'.

Once the concordance lines had been generated, the manual aspect of the investigation began, regarded in this study as a qualitative aspect of the analysis. This was the first step in the comparison of uses (functions) of the three verbs in the Expert and Student corpora, as part of addressing Research Question 3. The framework for the categorisation of the uses of each verb is explained in detail in Chapter 5 (see §5.8.3). This framework was used to code all the functions or categories of use of these verbs so that comparisons could be made within and across corpora. A very important function in this study was Concord's ability to provide evidence, such as the previous example, of the delexical use in the corpora of the three verbs, focusing as it does on how a word or phrase is used or behaves in context.

WST allows the researcher to insert codes into the concordance, which makes it possible to generate a graphical picture of the percentages of the various functions of a word in a corpus, using a program such as Microsoft Excel to do the calculations and generate the graphs or diagrams. In this step, each occurrence of each particular word-form was categorised according to the framework discussed below (§5.8.3).

Each function of the verb was classified and coded in Concord by assigning it a particular number in the 'set' column, where the coding is saved, allowing the researcher to rearrange the lines according to codes

(see Appendix E). Codes were assigned according to whether the verb was used as an auxiliary, as a lexical verb, as a semi-modal, as a phrasal verb or the category I was focusing on specifically, in a restricted or semi-restricted combination of delexical verb + noun/noun phrase. These categories are explained in more detail for each of the three verbs below (see §5.8.3.1–§5.8.3.3). These functions were then quantified using Microsoft Excel and pie charts were generated to indicate the proportion each function represented of the total occurrence of each verb. This element of the analysis dealt with the corpora as sub-corpora – Expert Lit and Law and Student Lit and Law.

Although this study does not purport to be a study of the entire grammatical workings and patterning of the three verbs in question, the functions and categories of use of these verbs were classified in order to isolate the focus of the study – the delexical uses of the verbs – from the rest. This process, involving the identification and explanation of each function of each of the three verbs, a search for these in the concordance of each word-form, and the coding of each word accordingly, has provided useful background information on students' overall use of these verbs compared to their use by expert writers. This is an aspect which has not been covered by any other researchers I could find, although Biber et al. (1999) provide information on all uses of these three verbs, which was invaluable to my study.

Finally, the lines containing core and pseudo delexical verb combinations featuring the three verbs in question were extracted.

Step 5 – analysis of delexical verbs

After the delexical combinations had been identified they were analysed in depth for frequency, appropriateness and differences across genres. This analysis meant establishing whether each example was a pseudo or a core delexical combination, as well as whether it was deviant in any way. If it was deviant, the errors were categorised and quantified.

Deviant MWUs were then identified and the errors they contained categorised and quantified. Once I had identified all the errors in these MWUs, two independent NS judges, colleagues from the Department of English Studies, both experienced, expert teachers of English language and literature, were asked to establish the degree of acceptability of combinations. The categories of acceptability were adapted from those used by Nesselhauf (2005): *Clearly Unacceptable *; Largely Unacceptable (*);* or *Marginally Acceptable ?*. I compared their ratings to my own, and arrived at a final, considered total for each category.

5.8.2 Classification of delexical MWUs

Based on studies by Algeo (1995) and Nesselhauf (2003, 2005) in particular, combinations were considered to be MWUs featuring a core delexical verb + noun/noun complement if they consisted of *HAVE, MAKE* or *TAKE* occurring

- with an eventive noun, where the verb carried little lexical weight or was semantically void or empty;
- where the noun carried the bulk of the semantic weight and where there was a verb, identical[24] in form, which could replace the whole combination, e.g. *let's have a look*, where *have a look* could be replaced with the verb *look*; *he took a walk* where *walk* could be replaced by the verb *walked,* since this definition allows for inflectional changes in tense and number.
- where the eventive noun was preceded by an indefinite article (a/an). For example,

study the literature if you do not **have a love for** reading per se. Some of you may (Exp Lit)[25]
purpose as a whole is to **make a comment** on the problem

An explanation of the classification of verb+noun combinations as delexical MWUs is provided in Chapter 4 (see §4.4.1). Combinations were considered pseudo delexical MWUs when they failed in some way to fulfil one or more of the requirements for core delexical MWUs:

- Where the replacement verb was not identical to the noun, but morphologically related as in *he made a decision* and *he decided* (Langer 2004:17), what Algeo (1995) terms 'affixation', but which is usually referred to as 'derivation'. This is the process of adding affixes to a word which typically changes the part of speech of the word, such as in the verb *decide* which can be changed to a noun, *decision,* by the addition of a suffix *-ion*. In contrast, as noted above, inflection changes a word grammatically, as in the inflection of the verb to indicate third-person singular in English (Richards and Schmidt 2002). Examples of this type from the Student corpus included

at we are going to examine. He **makes a very interesting observation** (Stud. Lit)

- Where there was 'a flaw in correspondence between the expanded predicate and a corresponding simple verb' (Algeo 1995:206), either through affixation, e.g. *make a decision – decide*, modulation (change of prosodic phonemes) and phonological modification (change of segmental phonemes), e.g. *make a prótest = protést, take a breath = breathe* (Algeo 1995:205), as in the example from the Student Lit corpus *The poet suggests that London **has a calming effect** on him,* or pluralisation, or deviations where the definite article is used instead of the indefinite article, or where the article is absent, or where there was no corresponding single-word verb in everyday use, e.g. *have a game*, *make an effort*, *have an affair*, or where there was only an equivalent non-cognate single-word verb, e.g. *take cover = hide* (Algeo 1995:206).

[24] 'Identical' is to be interpreted flexibly to accommodate necessary grammatical changes as driven by tense (and person).
[25] The source applies to the concordance line at the end of which it occurs and to any following lines. Where a different source is used in the same set of examples, a space is inserted and the rule applied accordingly. Corpora are referred to as Exp (Expert) or Stud (Student) Lit (i.e. Literature) or Law.

- Where the eventive noun was morphologically related to a simple verb, but the delexical MWU differed semantically from that verb, e.g. *make love ≠ to love*, *have a bite ≠ to bite* (Algeo 1995:206), as in the example *and watchful". It is because he* ***takes care*** *to note weaknesses* (Stud. Lit), where *take care ≠ care*.
- Where the corresponding simple verb was passive rather than active, e.g. *have a fright = be frightened* (Algeo 1995: 206).

In Langer's (2004) definition, the noun phrase in such combinations is not fixed and article and number may vary, and attributes (adjectives, intensifiers) may also be added. Negatives are allowed; and the phrase may also contain possessive pronouns. Such allowances are not made in Algeo's (1995) definition, and most of such MWUs would be classified as pseudo delexicals in my study.

I analysed all combinations carefully according to Algeo's framework discussed above, following the three rules for core delexical MWUs and the four rules for pseudo delexical MWUs in order to classify them. For instance, if the indefinite article was missing (for whatever reason) before the eventive noun in the target form, the combination was classified as a pseudo delexical MWU:

the first eight lines, the poet **makes use** of similes in order (Exp Lit)
the beginning and then ends up **making love** to her. He gets a

The second example here also illustrates the semantic difference between the MWU and the simple verb.

Once the delexical MWUs had been identified and categorised as core or pseudo with due consideration of their target forms, these combinations were investigated for possible deviations.

The following section discusses how the three verbs were categorised in terms of their functions. This is the focus of Research Question 3.

5.8.3 Analytical framework for analysis of *HAVE*, *MAKE* and *TAKE*

The frameworks for the analysis and categorisation of the uses of each of the verbs in question are presented in the following three sections, starting with *HAVE*.

5.8.3.1 HAVE

HAVE is a complex verb and has several functions. It is one of the three primary verbs (Biber et al. 1999) in English (the others are *BE* and *DO*), which means it can act as an auxiliary or a lexical (main) verb. It may also be used as an operator, a semi-modal and, according to the terminology used in my study, as a core or a pseudo delexical verb.

HAVE as auxiliary verb

HAVE is a primary auxiliary: it has inflections like lexical verbs, but these are frequently unstressed as contracted forms such as *'ve*. As a *primary auxiliary, HAVE* specifies the way in which the lexical verb, or the whole clause, is to be interpreted. The auxiliary *HAVE* is used to form the *perfect aspect*. The perfect aspect designates events or states taking place during a period leading up to the specified time (Biber et al. 1999:460):

Perfect aspect present tense (passive):

the norm. They, the Europeans, **have been** given this false (Stud Lit)

Perfect aspect past tense*:*

The deceased did not report that he **had recovered** his vehicle. (Stud Law)

The most common use of *HAVE* in the corpora in this study was as an auxiliary verb:

o Gatsby. Even though that may **have been** true, he neglected (Stud Lit)
away her purity. This may not **have happened** if the apartheid
air, putting things where they **have gone** wrong, for it is a
the speaker and the neighbour **have been closed**. As indicate
between him and Melanie should **have taught** him a lesson but
between Gatsby and Daisy, but **have** not yet **confronted**. He i

st what these other characters **have said**. He is definitely p (Exp Lit)

HAVE as operator

In general, *operators* are found in finite clauses only, and are used in special structures, particularly in independent interrogative clauses and clauses negated by ***not***. The operator is realised in the following ways: by the first auxiliary in the verb phrase: ***Are*** *you joking?* and also by the insertion of the auxiliary **do**, as in these examples with *HAVE*:

insight into his own character? **Does** he **have** any respect for Soraya's? (Stud Lit)
is suddenly terminated? **Does** Lurie **have** the potential for sustaining;

and by the copular verb ***be*** and, less commonly, transitive ***have****.* The examples below are provided from Biber et al. (1999:134), as there were none in my corpora:

Have you any money?
I **have** a lot of feeling – right?
I **haven't** any money.

The operator can also be used to 'underline the truth of a positive statement' and 'this is stressed. If not, the auxiliary **do** is inserted' (Biber et al. 1999: 134):

of the allusion to the hero Oedipus? **Does it have** to do with the manner in which Lurie embarks (Exp Lit)
an insight into his own character? **Does he have** any respect for Soraya's feelings? If yes,
is suddenly terminated? **Does Lurie have** the potential for sustaining certain relation

There were very few examples of *HAVE* as operator in the corpora.

HAVE as semi-modal

HAVE can be used as a *semi-modal* verb, which is a use related to modal auxiliary verbs (*can, could, may, might, must, shall should, will, would*). Modal auxiliary verbs differ from other verbs in that they have no non-finite forms. These modal auxiliaries express a wide range of meanings, having to do with concepts such as ability, permission, necessity and obligation. Although they can convey meanings that relate to time differences (e.g. *can* vs *could*), the differences among them relate primarily to modality rather than to tense. The verbs *dare (to), need (to), ought to,* and *used to* are on the borderline between auxiliaries and lexical verbs and can be regarded as marginal auxiliaries. These vary with respect to *do*-insertion. *HAVE,* which is related in meaning to modal auxiliaries, occurs as part of a multiword verb as in *have to, (had) better, (have) got to* and in this use is called a semi-modal (Biber et al. 1999:73). Biber et al. (1999:484) describe these semi-modals as 'fixed idiomatic expressions with functions similar to those of modals'. These expressions are commonly contracted, as in '*d better*. Expressions with *HAVE* as a semi-modal, as in *have to,* express obligation or necessity:

oman is a burden on which some **have to** overcome it and some (Stud Lit)
urs make good friendship. They **have to** close the gaps and bo
abamukuru is often praised for **having to** take care of both t
bamukuru in telling submission **have to** be married women whil
ighed again. The two recording **have to** agree and if not, it
ect, feel for each other. They **have to** help without limits a
d up by what he sees, it would **have to** be a dull person. The
e neighbours, at one time they **have to** meet and let each oth

HAVE as lexical verb

HAVE as a transitive lexical verb occurs as commonly as the most frequent lexical verbs in English. Biber et al. (1999:429) found that across their four registers of conversation, fiction, academic prose and news, *HAVE* was most common in conversation and least common in academic prose. However, within academic prose, *HAVE* was more common than any of the lexical verbs. The main verb *HAVE* can be used with various meanings indicating different logical relations, including physical possession, family connection, food consumption, states of existence, linking a person to some abstract quality, marking causation. Below is a sample from the concordance lines for *hav** as a lexical verb in the Student Lit Corpus:

implies that as long as they **have the wall**, they will be (Student Lit)
at they all go to the City and **have some fun**. They booked a
ne that these neighbours would **have an idea** on how these ten
not see light. I think if you **have gift** from God you will n
and absence of light; his eyes **have no light** to even create
is not fair as if it does not **have anything** fair to human e
hat he does makes him free **and have a clear conscience**. Yes,

rced to use their own toilets, **have their own doors** that the
ll'. The river is described as **having its own will** (line 12)
talks to Agnes. He is said to **have shadowless eyes** and cada

HAVE as delexical verb

The main focus of this study was those occurrences where *HAVE* was used delexically (Altenberg and Granger 2001:174ff) with an eventive noun. Algeo (1995:204; cf. Quirk et al. 1985:750) explains an 'eventive' object as 'deverbal and [...] the exponent of the meaning of its verbal correlate'. There are two types of delexical construction identified in this framework, which I have called 'core' and 'pseudo' delexical MWUs respectively, borrowing the terms from Algeo (1995).

HAVE as core delexical verb

Biber et al. (1999:1026) explain that verbs like *HAVE*, *MAKE* and *TAKE* are often found in combination with a noun phrase, forming 'relatively idiomatic expressions'. These relatively idiomatic combinations can often be replaced by a single-word verb with essentially no loss of meaning. In this study, the kind of combinations which Biber et al. (1999) regard as 'idiomatic', relatively or completely (see Chapter 4, §4.4.1), I have coded as core delexical verbs. Below are examples from the Expert Lit and the Student Lit corpora:

the literature if you do not **have a love** for reading per se. Some (Exp Lit)
in the examination, to **have a grasp** of the basic ideas, and aspects
en kept secret for so long? He **has a newfound respect** and ap

ourage and love of animals and **has a respect** for her and her (Stud Lit)
touches the speaker deeply and **has a spiritual impact**. 'The

HAVE as pseudo delexical verb

Where the eventive noun or expanded predicate and the corresponding simple verb are not identical either through affixation, modulation and phonological modification or pluralisation (although grammatical changes to accommodate tense and person are allowed), or where the definite article is used instead of the indefinite article, or where the article is absent, these are termed pseudo delexical combinations (after Algeo 1995). The example below

present life? You may not **have much respect** for a character (Exp Lit)

explains one of the difficulties of analysis some of these MWUs presented. This example contrasts with the one for the core delexicals above in that it does not take an article, partly because of *much*, which requires a non-count interpretation. Thus, quantification, an aspect of the grammar, can determine whether a structure is regarded as pseudo delexical or not. In the examples below, the combinations can be replaced by single-word verbs but these are not identical in form to the eventive noun:

dd looking guy. He is tall and **has a ghostly appearance**; his (Stud Lit)
at the narrator is speaking of **has no comparison** in the natu
and realises **the effect Uriah has** on her father. Uriah Heep
ncentration The word '**majesty' has a connotation** of our almi
explains clearly that the man **has no feelings** and is always
tive and thoughtful person who **has logical thinking** when awk
tive if someone ignore him. He **has an aggressive behaviour** w

5.8.3.2 MAKE

MAKE is a lexical verb only, never acting as an auxiliary and never used intransitively (that is, without an object). *MAKE* can be classified as an activity verb according to its core meaning, which is to create something, but it may also be causative in meaning – *to make something happen*, *to make (force) someone to do something*. As an activity verb, Biber et al. (1999:367) found that it occurred more commonly across all four registers (and over 1000 times per million words in the academic register) than any other of the activity verbs – the list was topped by the 'light' verbs *GET*, *GO*, *GIVE* and *TAKE*. *MAKE*, *TAKE* and *GIVE* occur often with the progressive aspect (i.e. more than 10 times per million words) (Biber et al. 1999:471).

In addition to functioning as a lexical verb, *MAKE* can function as a phrasal verb, a prepositional verb and a phrasal prepositional verb. It can also be used delexically as a core or pseudo delexical verb. Biber et al. (1999:403) observe that *MAKE* is used in relatively idiomatic units that function like single verbs, that is, phrasal verbs, prepositional verbs and phrasal-prepositional verbs.

MAKE as phrasal verb

A phrasal verb is made up of a verb (in this case *MAKE*) with an adverbial particle, e.g. *make up, make out*. The core meaning of the adverbial particle indicates *location* or *direction*. There were only a few examples of phrasal verbs in the corpora in this study, including:

at it was forced upon him'. He **made up** people to be more imp (Stud Lit)
ous rules and regulations that **make up** the apartheid regime.

rds and actions the playwright **made up**, that she chose becau (Exp Lit)
something that is, literally **'made-up**', which is what 'fict
that a few fleeting encounters **make up** a substantial relation

a because those immigrates are **making out** our country overcr (Stud Law)

MAKE as prepositional verb

Prepositional verbs consist of a verb followed by a preposition, such as in *make for; made of, made from:*

t to rebuild the roof of house **made of** grass. During winter (Stud Lit)
ner. They drunk Mr Smit's brew **made of** strawberries. This wa

Philip, my infant tongue could **make of** both names nothing lo (Exp Lit)
cut out all the 'frills' which **make for** an attractive paragraph

ibery, or where allowances are **made for** others based on many (Stud Law)

MAKE as phrasal prepositional verb

Phrasal prepositional verbs contain an adverbial particle as well as a preposition, for example, *make up for*, and *be made up of*. There were only a few occurrences of *MAKE* as a phrasal-prepositional verb in these corpora:

the body of the essay should be **made up of** a series of linked (Exp Lit)
resent tense, its main verb is **made up of** a helping verb ('a

climax. The English sonnet is **made up of** an octave and a se(Stud Lit)

he many societies of the world **made up of** the good and the b (Stud Law)

MAKE as single-word lexical verb

As noted above, *MAKE* can be classified as an activity verb according to its core meaning, but it may also be causative in meaning. *MAKE* belongs to the group of activity verbs that occurs with transitive patterns only.

with the statement saying "And **makes gaps** even two can pass (Stud Lit)
is kind and considerate. If he **makes a partner** or rather whe
bours". That means good advice **make good son**. Definition: It
nd believing that "good fences **make good neighbours**". When t
s able to soften her heart and **make her believe** that he was
edient is, she did not want to **make her father angry**. She di
of male dominance (her father) **makes her more** question the w
Niki's action is her desire **to make her baby** more acceptable
ple, appeals to our senses and **makes us imagine** that we are

MAKE as core delexical

Where relatively idiomatic combinations with *MAKE* could be replaced by a single-word verb with essentially no loss of meaning, these were classified as core delexical MWUs. Together with *HAVE* and *TAKE*, *MAKE* is 'particularly productive when combining with a following noun phrase to form relatively idiomatic expressions' (Biber et al. 1999:1026–1028). These expressions form what Biber et al. (1999:1026) refer to as a 'cline of idiomaticity'. At one end are expressions which are clearly idiomatic, such as *make a killing*. At the other end of the cline are expressions which 'retain the core meaning' of these verbs: *he made a sandwich* (Biber et al. 1999:1027). Between these extremes, there are a number of expressions where the meaning of the individual words is to some degree retained but the whole expression has a more idiomatic meaning: *make a deal*, *make a statement*. These can in many cases be replaced by a single verb which is identical in form (although allowing for changes necessitated by tense or number) to the (deverbal) noun in the combination (Biber et al. 1999:1026–1027). In my study, only these combinations

are categorised as core delexical MWUs. Below are examples of core delexical combinations with *MAKE* from the corpora:

in Africa because they want **to make a profit**. He sees the de (Stud Law)
ermore, the person in (a) must **make an offer** to the person I

in their imagination. The poet **makes a contrast** between the (Stud Lit)
e. There again Tom should **have made a plan** to protect Jay. I

ng of your work. It is wise **to make a list** of words you freq (Exp Lit)
'solves' his sexual problem by **making a weekly visit** to Sora
tet's purpose as a whole is **to make a comment** on the problem

MAKE as pseudo delexical

MAKE as a pseudo delexical verb is classified according to the same criteria as discussed for *HAVE* above. Biber et al. (1999:1027) provide a list of noun phrases which combine with *MAKE* in their corpus of conversation or fiction: *an appointment, arrangements, a/the bed, a (clean) break, a decision, a/any/no difference, an/every effort, eyes, a fool of (one)self, fun, a fortune, a bad/good impression, a killing, a living, love, matters (worse), a mistake, money, a move, a noise, a nuisance, a pest of oneself, plans, a point, a sound, a speech, a statement, time, work*.

Noun phrases combining with *MAKE* in news and academic prose that are listed by Biber et al. (1999:1027) and which would be classified as pseudo delexicals in this study include: *amends, assumptions, choices, comparisons, demands, headlines, his/her appearance, her/his comeback, her/his debut, history, judgements, predictions, provision, recommendations, reference, use of*. If these are used in the plural form idiomatically, such as *make amends*, I regarded these as pseudo delexical combinations. Where the nouns were derivations or affixed forms and could not be replaced by a single-word verb with the same form (other than changes necessitated by tense and person) and meaning, these were also classed as pseudo delexical combinations.

Biber et al. (1999:1028) observe that although it is generally believed that formulaic language occurs predominantly in conversation, their corpus showed that idiomatic phrases including *HAVE*, *MAKE* and *TAKE* are much more common in written registers, and several such expressions are common only in news reportage and/or academic prose. They found that, in particular, idiomatic phrases with *MAKE* and *TAKE* occurred frequently in written exposition. Pseudo delexical uses of *MAKE* in the corpora in my study included:

ty which has not yet woken. He **makes reference** to the 'smoke (Stud Lit)
nd did have the opportunity to **make independent choices** and
without him. She is willing to **make sacrifices** in order to r
en Lurie is cross-examined, he **makes some startling statements**
he had done and an attempt to **make reparation**, something he

ot be a fair **choice to have to make**, but it is a choice none
to the autobiographical, to **make explicit the link** between
text, yourself, and the author **make connections** to yourself,

5.8.3.3 TAKE

TAKE belongs to the group of activity verbs (like *MAKE and GIVE*), verbs which 'mainly denote actions and events that could be associated with choice, and so take a subject with the semantic role of agent' (Biber et al. 1999:361). *TAKE* is the fifth most common of the 12 lexical verbs which occur over 1000 times per one million words across Biber et al.'s (1999:373) four registers and it is one of six verbs in this group which are from the activity semantic domain. These verbs are most common in conversation and least common in academic prose, accounting for only 11% of lexical verbs in this register (Biber et al. 1999:373). The reason for this is that academic prose mostly

> reports relationships between entities – both concrete and abstract – using simple statements of existence/relationship or occurrence. Academic prose reports relatively few physical, mental, or communication activities – and when such activities are reported, they are often attributed to some inanimate entity as subject of the verb (Biber et al. 1999:372).

In academic prose *TAKE*, like *MAKE*, occurs commonly with an inanimate subject that is in some way instrumental to the meaning of the verb. In addition, instead of describing an activity or process, the verbs in these uses tend to describe static situations or relationships, becoming existence verbs, as in the following examples:

*'[Testing] usually **takes** the following three steps*
*Social [science], [religion], and the [arts], **make** contributions'* (Biber et al. 1999:379).

TAKE is one of those activity verbs which occur with transitive patterns only, although exceptions do occur in colloquial speech: e.g. *it won't take* … . It can also occur as a phrasal verb, as a prepositional verb and as a phrasal prepositional verb. The occurrences of *TAKE* in the corpora in this study were coded according to the following functions or classifications:

TAKE as phrasal verb

Phrasal transitive verbs with *TAKE* include *take up* (occurs over 20 times in fiction and academic prose and over 40 times per million words in news reportage) (Biber et al. 1999); *take on* (over 20 times per one million words in fiction and academic prose, and over 40 times in news); *take off* (over 40 time per million words in fiction); *take over* (over 40 times per million words in news) (Biber et al. 1999:410). *Take up* and *take on* are more common in writing than in conversation, which is contrary to the norm for phrasal verbs (Biber et al. 1999:412). *TAKE*, as one of the most common lexical verbs, is also most productive in combining with adverbial particles to form phrasal verbs. This is because *TAKE* is 'unusually polysemous' (Biber et al. 1999:412), meaning that it is able to combine with several adverbial particles: *take* + *apart*,

back, *down*, *in*, *off*, *on*, *out*, *over*, *up*. In fact, *TAKE* forms seven common phrasal verbs (i.e. over ten per million), and nine phrasal verbs with *TAKE* are listed in the Longman Dictionary of Contemporary English (LDOCE). The following are examples of *TAKE* functioning as a phrasal verb, taken from the Student Lit corpus:

wrong deed was for Antonio to **take away** Shylock's way of li (Stud Lit)
Niki when she tried to scream. **Taking off** her underwear and
hat his mental health has also **taken on** a dark state of mind
lised person who is seen to be **taken over** by greed and the p

TAKE as prepositional verb

TAKE can also be a prepositional verb when it takes a prepositional object, e.g. *take NP as, take NP for, take NP from, take NP to, take by:*

purity. The word Excelsior is **taken from** Latin word excelsi (Stud Lit)
by her father These injustices **take us to** a point where Niki
ndent on Babamukuru because he **takes her into** his home, pays
e is a metaphor, the river was **taken as** a human being specif
David could suddenly see what **taking by force** meant. Lucy e
guru was a sacrifice women who **take Tambu as** her child and t

TAKE as phrasal prepositional verb

TAKE, like *MAKE*, occurs as a phrasal prepositional verb, with an adverbial particle as well as a preposition, e.g. *take NP away from, take NP out of*:

ther's shack. She has plans of **taking her mother out of** that (Stud Lit)
looked down upon and finally **taken away from** him. The weal
ght in shining armour who will **take her away from** all her mi

TAKE as single-word lexical verb

TAKE occurred in my corpora with transitive patterns only:

m is fast and the rhyme scheme **takes a division** of two group (Stud Lit)
n life capacity of renewal. He **take a job** at the animal shel
lamed for his actions. He does **take a degree** of responsibilies
resist temptation and chose to **take their relationships** with
nhuman, bigoted people who had **taken power**, wealth, educatio
he is in Cape Town, he is still **taking prostitutes** though he
change. She realises she must **take her own turn** and be allo
s his. (Never mind that he was **taking several mistresses**.) T

TAKE as core delexical

TAKE, like *MAKE* and *HAVE*, is what Biber et al. (1999:428) call a semantically light verb, meaning that it can combine with noun phrases to form a set verbal expression. In the case of *TAKE*, these include clear

idiomatic expressions such as *take time*. Then there are those expressions that keep the core meaning of this word (Biber et al. 1999:1027): *you can **take** a snack in your pocket*. Between these expressions are the 'relatively idiomatic expressions' such as *take a walk* where the meanings of the individual words are to some extent preserved, but the expression as a whole also assumes a more idiomatic meaning. The commonest of these combinations in academic writing with *TAKE* are: *take place; take part; take advantage (of); take (the) form (of)* (ibid: 1028), all of which would be classified as pseudo delexical verbs in my study. Biber et al. (1999:1028) found that, contrary to popular perception, these idiomatic phrases are far more common in written registers than in conversation, and several of them occur only in academic prose and news reportage. Below are examples of core delexical combinations from the corpora:

eedy. Greed is from the devil; **take a look** of the following (Stud Law)
l problem but worldwide. If we **take a look** around the world
do believe that if corruption **take a lead** it will takes us
s are presented to the people. **Take a look** at the top manage

allousness, she feel strong to **take a step** forward for her (Stud Lit)
through the window while Lucy **took a shower**. David saw him

TAKE as pseudo delexical

As noted above, only some of the expressions Biber et al. (1999:1027) identify as 'relatively idiomatic' are categorised as core delexical verbs in my study: those which cannot, for the same reasons as discussed for *HAVE* and *MAKE* above, belong in the group of pseudo delexical MWUs. Examples of pseudo delexical MWUs with *TAKE* from the corpora include:

and Madam Cronje's husband also **take part**. The eventual result (Stud Lit)
example where changes in rhythm **takes place**. The full stop at
he was interested to see Agnes **taking the step** because after
skyscrapers, that people do not **take the time** to appreciate t
is love. Because he decided to **take the fall** for the accident

These combinations were then extracted and the errors were categorised and analysed. This study makes a distinction between the terms 'deviation' and 'error': 'deviation' refers to those MWUs which were in some way problematic, while 'errors' refers to each way in which such MWUs were deviant. The following section provides a detailed explanation of the framework for the analysis and categorisation of these errors.

5.8.4 Classification of errors

Errors in deviant MWUs were classified and are discussed here in detail with reference to examples from the concordance lines. The classification is based loosely on Nesselhauf's 'Types of mistakes in collocation' (Nesselhauf 2003:232) but with sub-categorisations included, and with the category of adjective (ADJ)

added. Thus, seven main categories of error were identified, namely, those concerning chiefly the verb (V), the determiner (D), the noun (N), the preposition (P), the adjective (ADJ), those involving structure (S) and those involving stretched verb constructions (SVC). These are discussed in more detail below (and the full list of deviations is provided in Appendix F):

- ADJ – adjective. Examples included collocation errors as in the example below where the adjective is not a fully appropriate collocate with *living* (ADJ – coll):

    1. d with drugs. Drug dealers are **making a wealthy living** out o (Stud Law)

- D – determiner. Deviations in the determiner include errors in the article (D – article), as in (1) missing article (D – article missing), where the indefinite article *a* is missing after *has* (*has a life*); (2) incorrect article (D – article incorrect), where the wider context revealed that *a* should be replaced by the definite article *the* (*has the connotation*), or (3) present but inappropriate article (D – article inappropriate). In the example below, besides the error in the verb (concord), the article is present but inappropriate in *have the disgrace*:

    1. They make the city of London **has life** like a living being. (Stud Lit)
    2. title The Madonna of Excelsior **has a connotation** of the women
    3. Melanie's father. Lurie says he **have the disgrace** for his who

  Deviations in the determiner also include errors to do with the pronoun (D – pronoun): (1) missing pronoun, in this case the possessive pronoun *her*; (2) incorrect pronoun, present but unacceptable, in this case *his* should be replaced by *her*:

    1. alerting Agnes to stand and **make voice** be heard. The narr (Stud Lit)
    2. a councillor meanwhile **she is making *his* way** to the top by

  Finally, deviations in the determiner included errors to do with the demonstrative (D – demonstrative): (1) incorrect demonstrative, in this case where *this* should be replaced by *these*, to agree in number with the plural *notes*:

    1. they come with this technology of **making *this* notes** this lead to (Stud Law)

- N – noun. Errors in the noun concerned number: (1) plural where singular was required and vice versa, as in the example below where *mistake* should be plural to agree with '*a lot of*':

    1. got in this world. David **made a lot of mistake** in life (Stud Lit)

- P – preposition (P). Errors also occurred in prepositions: present though incorrect, as in the examples below, where *off* should be replaced with *of* in (1) and *to* should be *for* in (2). These prepositions are determined by the MWU in each case and so these examples are identified as errors:

    1. at she and her family would be **taken care *off*** if she does no (Stud Lit)

2. oom. He has changed because **he has sympathy *to*** his daughter.

- S – structure (S). These were errors of syntax where (1) the structure of the whole expression was incorrect. In this case interpretation is not straightforward. The italicised expression could be replaced with *If a bribe is what it takes?* or *If it takes a bribe (to somebody)*:

1. o anything achieve that. *If it* ***take to bribe*** *somebody* in order (Stud Law)

- SVC (SVC) – stretched verb construction. This is Nesselhauf's (2005) term for what are here referred to as delexical MWUs. In my study, these deviations comprise cases where a simple verb would be more appropriate than an SVC (i.e. MWU):

1. this fellow Uriah. He does not **have any trust** *towards* this (Stud Lit)

In this example, a single verb *trust* would have been preferable to the SVC *have any trust towards*

- V – verb. Deviations involving the verb included: (1) tense (V – tense), in most cases the overuse of the progressive aspect present tense (*to be having buzzing activities* instead of *to have*); (2) concord (V – concord), such as subject-verb agreement; (3) errors in collocation (V – coll), such as *make corruption;* and (4) errors in choice of verb (V – word).

In example (1) below there is an error in the tense of the verb, where the progressive aspect present tense should have been replaced by the present tense *has*:

1. lines sestet where the poet is **having a solution** *of* his problem (Stud Lit)

Example (2), in addition to other errors, contains an error of concord, where *have* should be replaced by *has*:

2. and the people was dying. He **have no hope** *of* the universe (Stud Lit)

The next type of verb error (3) was problematic collocation between verb and noun in the MWU:

3. today's newspaper. Immigration **has had a great contribution** (Stud Law)

In this example, the student has collocated *has* incorrectly with *contribution*. *Contribution* collocates with *make* (*make a contribution*).

In Example (4), the choice of verb is incorrect (*heed* should have been selected rather than *hid*). This may simply be a spelling error, however, as there is no distinction between long and short vowels such as *heed* and *hid* in the pronunciation of many BSAE speakers.

4. away, advice that he does not **take hid** of. Gatsby could also (Stud Lit)

In many cases there was more than one error in the MWU, which sometimes had the cumulative effect of making interpretation difficult. In all cases, each error was coded and explained. Where, occasionally, one error caused another, such as in the example

awful. Yes I agree some of them **have connection** with our guys (Stud Law)

where there is an error in the determiner (D – missing indefinite article after *have*) and an error in the noun (N – *connection* should be plural), only one error was counted, in this case the lack of a determiner. Choices in such cases were sometimes not easy, but here the rationale had to do with the key role that the determiner plays in the definition of the focal structures of this study, the delexicals.

In other cases, such as in the example below, errors occurred close to, but outside the MWU. Such errors were not included in the analysis or discussion:

ntion that *their* are coming **to make better living**. But not a (Stud. Law)

In this example, the possessive pronoun *their* has been inappropriately used instead of *they* but as it occurs outside *to make a better living* it is not included in the error count.

Once all deviations had been classified, frequencies could be computed and compared across corpora (see §5.3).

5.9 PHASE 3: THE RELATIONSHIPS BETWEEN STUDENTS' VOCABULARY SIZE, PRODUCTION OF MWUS AND ACADEMIC PERFORMANCE

This part of the analysis was completed using SPSS and WordSmith Tools. A smaller group of 60 students was selected for this part of the study, using a form of non-probability sampling (Dörnyei 2007:98), known as quota sampling (see §5.4.1). In this type of sampling, the size of the sample selected within a particular sampling framework is predefined. Using examination scores as the organising variable, ten students whose scores fell closest to these percentiles were drawn from each of the 25th, 50th and 75th percentiles for both the Literature and Law groups. The texts of these 60 students were extracted from the Student sub-corpora and concordance lines were generated for all word-forms of *HAVE, MAKE* and *TAKE*. The delexical MWUs in these concordance lines were extracted, deviant MWUs were identified and errors categorised. In addressing Research Question 5, scores on the 5000-word level, the UWL and the total vocabulary scores of these 60 students (30 from Literature and 30 from Law) were compared to the number of MWUs and errors. Correlations were performed in this phase of the analysis. In addressing Research Question 6, I

investigated students MWUs and errors in relation to their academic performance. The results of these analyses are presented and discussed in detail in Chapter 6.

5.10 JUSTIFICATION OF TECHNIQUES

As this chapter has indicated, each phase employed several methods which have been well documented in other studies, with good results (e.g. Wang and Shaw 2008). The study makes use of quantification and statistical analysis in the first and third phases of the analysis, and of quantitative and qualitative analysis of texts in the second phase, again following in the tradition of similar studies (Altenberg and Granger 2001; Boers et al. 2006; Cobb 2003; De Cock, Granger, Leech and McEnery 1998; Granger 1998b; Hasselgren 1994; Laufer and Waldman 2011; Lee and Chen 2009; Nesselhauf 2003, 2004, 2005).

The second phase of the study follows a recognised methodological trend in the field of corpus linguistics and has a place as a corpus-driven study dealing with MWUs. The use in this study of a smaller corpus, in this case of writing by learners, is also well supported by studies as noted above (§2.3.2). Scholars such as Flowerdew (2001) Nesselhauf (2005) and Shirato and Stapleton (2007) advocate the use of smaller, often self-compiled learner corpora if learners' difficulties are to be insightfully addressed. In South Africa in recent years there has been an increase in studies of both learner (Van Rooy 2006) and non-learner varieties of language (De Klerk 2002; Pienaar and De Klerk 2009) and several smaller corpora of language have been collected in this regard.

The focus of Phase 2 on word combinations featuring high-frequency delexical verbs is also well supported by research studies which reinforce the assumption explored here that these seemingly unimportant words often form part of problematic and recurrent word combinations in learner writing (Altenberg and Granger 2001; Laufer and Waldman 2011; Lee and Chen 2009:154, Nesselhauf 2005; Wang and Shaw 2008). This study has included a further dimension to the study of these words, however, in its analysis and comparison of the distribution of these words in the Expert and Student corpora and in the Lit and Law corpora. In addition to the manual analysis of the concordance lines and the categorisation of errors, log-likelihood calculations were the chief statistical measure used in this part of the analysis, used to determine the difference in frequency of use of MWUs in the different corpora.

5.11 ETHICAL CONSIDERATIONS

The ethical considerations are touched on in Chapter 1. As I tested the vocabulary size of individuals, there were ethical considerations regarding anonymity and confidentiality. No names appeared anywhere on the questionnaire or test scripts, or on the examination scripts. Students were assured from the outset that no

names would be used, and that student numbers would at no point be linked to student names; student numbers were used only to distinguish one file from another in the Student corpus and to link vocabulary test scores to examination scores, but these files or scores were not linked to individual students in the discussion. The students' examination and VLT scores were linked and used in order to determine the proficiency levels of the sample as a whole and, again, individual students were not named anywhere in the discussion of the results. Smaller samples of texts were extracted from the corpus in Phase 3, but these were linked neither to individual students' names nor student numbers at any point in the discussion. Students were consulted at the outset, through a letter attached to the questionnaire and test which was sent to them by post (see Appendix A), and were asked to give permission for their test scores and examination results, as well as their examination scripts, to be used in the study. They were assured that they would be free to withdraw their consent at any time should they so wish; none of the students did so, however. Before the research was set in motion, the study proposal was passed by the University's Ethics Committee and Dean of the College of Humanities at the university also gave her permission for the study to proceed (see Appendix B).

As far as the Expert corpus is concerned, I was at one point several years ago involved in the tuition of both the Literature and the Law courses. This meant that I contributed to some of the study material and examination papers at the time. As these texts were all written before the study was conceptualised, however, and since I no longer teach these courses and do not set any of their examination papers, my prior involvement had no influence on the content of either Expert or Student corpus and would not have resulted in bias.

5.12 CONCLUSION

This chapter has explained the methodological 'nuts and bolts' in each of the three phases of the study – the practicalities and key problems involved in gathering and analysing the data in an effort to provide answers to the research questions. Problems included the difficulties involved in getting tests and questionnaires back from a sample of students, of whom the majority study at a distance from the university campus. This chapter has also provided a description of the sample from which the data were collected and has explained the methods used to analyse these data. In Phases 1 and 3 this analysis was quantitative and involved the use of SPSS to perform the correlations and other statistical procedures (see Appendix G for a sample of SPSS output). In Phase 2, WordSmith Tools, most particularly the Concord tool, was used to explore the corpora. Statistical tests such as log-likelihood calculations were also used in this phase. This chapter has also laid out the frameworks for the analysis of the concordance lines for each verb, and for the classification of the delexical MWUs and the errors which occurred in deviant combinations. Finally, the ethical considerations were discussed.

Chapter 6

Analysis and Discussion of Findings

6.1 INTRODUCTION

This chapter presents the findings of the analysis of the data collected in this study. The findings of each of the three phases and for each research question (RQ) are presented and discussed in turn. To recap, in Phase 1, students' vocabulary size was measured using a version of the VLT (RQ1). The relationship between vocabulary size and course, gender, age and home language examination scores was investigated. Phase 2 of the study comprised an investigation of the use of three high-frequency verbs in terms of their functions by student writers and expert writers, using WordSmith Tools (WST). The use of these MWUs by students, and the errors they made in their production, was investigated and compared to their use in the Expert corpora. In Phase 3, the relationship between students' vocabulary size and their ability to produce appropriate and idiomatic MWUs containing these high-frequency delexical verbs was investigated, as well as the link between the students' use of these MWUs and their academic performance.

6.2 PHASE 1 SIZE OF PRODUCTIVE VOCABULARY (RQ1) AND ITS RELATIONSHIP TO ACADEMIC PERFORMANCE (RQ2)

Phase 1 is quantitative in design as it measures the size of students' productive vocabulary using the VLT and relates their scores on this test to their examination scores. The first part of the analysis in this phase addresses the size of students' productive vocabulary knowledge. This question is broken down into four sub-questions, which are addressed in turn. The second part of this phase addresses the relationship between the size of students' productive vocabulary and their academic performance. Statistical analysis in this phase was carried out using SPSS version 19 (see §5.7.3).

6.2.1 Questionnaire data

The first step in the analysis was to examine the biographical data provided in the questionnaires and obtained from registration records. In this study I focused on only a selection of biographical aspects, namely, course of study, gender, age, and home language, though data on ethnicity is sometimes also referred to.

6.2.1.1 Students' course of study and gender

Table 6.1 indicates the make-up of the students who returned completed questionnaires, according to their course of study and gender:

	Course of study		
	Literature	**Law**	**Total**
Male	28	95	123
Female	111	64	175
Total	139	159	298

Table 6.1: Students' gender according to course of study

This reflects a split of approximately 20:80 between male and female students in the Literature course, and 60:40 in the Law course. This suggests that women may be more likely to favour the arts than men, who may be more vocationally driven than women students. This is discussed further in §6.2.2.

6.2.1.2 Student age

Students varied widely in age: a third of the 284 students who provided their age on the questionnaire (35.2%) were 25 years of age or younger, and 10.2% of the sample was younger than 21, as shown in Table 6.2 below:

Age	**Course type**		
	Literature	**Law**	**Total**
< 20–29	69	69	138
30–39	36	50	86
40–49	16	17	33
> 50	14	13	27
Total	135	149	284

Table 6.2: Student age bands

Almost half the students in the sample were under the age of 30 (51.1% of the Literature students and 46.3% of the Law students, that is), certainly what one would expect from typical university students. However, these are distance education students and it is perhaps surprising that so many of them were younger: 10.2% were under the age of twenty and 25.0% between the ages of 20 and 25. The students in the 30 to 39 year age group are probably more typical of distance education students, possibly starting their university studies some years after leaving school. At the other end of the scale, 'mature' students of 40 years or older made up about a fifth of the sample (21.1%), with almost half of this group being over the age of 50. The oldest student in the sample was 68 years old and the youngest, 18. Almost a third of the sample (30.3%) was aged between 30 and 39 years and almost 80% of students in both groups were under

the age of 40 (77.7% in the Literature group and 79.8% in the Law group). This is an indication of the degree to which institutions such as this cater for individuals who, through various life circumstances, may have had to defer higher education for some years after leaving school, but for this very reason it is also somewhat surprising that there was not a larger group of older students, over the age of 40.

There is a fairly even spread of students across the age bands in the two courses, except in the 30 to 39 group, where more Law than Literature students were represented. In both genres, there is an almost 50–50 split between students under 30 years of age and those above 30 (51.1% and 48.9% respectively in the Literature course and 46.3% and 53.7% in the Law course).

6.2.1.3 Student home language and ethnic group

South Africa has 11 official languages and it was anticipated that the students in this sample would be representative of the national profile which is, according to data from the 2011 census, as illustrated in Table 6.3:

Language	Percentage of speakers*
Afrikaans	13.5
English	9.6
IsiNdebele	2.1
IsiXhosa	16.0
IsiZulu	22.7
Sepedi	9.1
Sesotho	7.6
Setswana	8.0
SiSwati	2.5
Tshivenda	2.4
Xitsonga	4.5
Other	1.6

** Spoken as a home language*

Source: Census 2011

Table 6.3: Percentage of speakers of South African languages 2011

The language profile of students in the present study was as follows (arranged in descending order of frequency):

Language	Number of speakers*	Percentage of total
English	119	39.9
Afrikaans	37	12.4
isiZulu	30	10.1
Sepedi	23	7.7
Setswana	21	7.0
isiXhosa	19	6.4
Other	14	4.7
Sesotho	8	2.7
Xitsonga	8	2.7
SiSwati	3	1.0
Tshivenda	3	1.0
isiNdebele	2	0.7

** Spoken as a home language*

Table 6.4: Languages spoken by students in the study

There was a high percentage of English speakers (39.9%, as opposed to the national average of 8.2%), and this is to be expected as it is the language of learning and teaching (LoLT) of the majority of schools in the country, and of this university.

The home language according to ethnic group is reflected in Table 6.5 below. Although this information is included here merely to allow further insight into the language background of students in the sample, it is a requirement of the South African government that students state their race on their registration forms. The main purpose behind this requirement is the redress of past inequalities and injustices through affirmative action. The official groups used by the government are white, black, coloured and Asian.

Language	Asian	Black	Coloured	White	Total	%
English	21	29	7	62	119	41.4
Afrikaans	0	1	7	29	37	12.8
isiZulu	0	30	0	0	30	10.4
Sepedi	0	23	0	0	23	8.0
Setswana	0	21	0	0	21	7.3
isiXhosa	0	19	0	0	19	6.6
Other	0	11	0	3	14	4.8
Sesotho	0	8	0	0	8	2.7
Xitsonga	0	8	0	0	8	2.7
SiSwati	0	3	0	0	3	1.0
Tshivenda	0	3	0	0	3	1.0
isiNdebele	0	2	0	0	2	0.6
Total	21	158	14	94	287	

Table 6.5: Languages spoken by student sample according to ethnic group

As noted earlier (§5.6.2), there is evidence from research that suggests that black students such as those in this study, who stated in their questionnaires that their home language was English, are in the main likely to be multilingual (Coetzee-Van Rooy 2012, 2014). Such students would be speakers of an African language as home language, but would frequently regard English as their strongest language, and the language used in the domain of education; the listing of it as their 'home language' could in fact be an indication that they wished to indicate that they regarded this as their strongest language in terms of study at university. According to Coetzee-Van Rooy (2012:112), this is evidence of a 'flourishing functional multilingualism' where the home language is used more for social communication and English is the language of education.

Asian students were probably mostly of South African Indian descent and thus first language speakers of English, albeit perhaps the variety of South African Indian English (SAIE, identified by Pienaar and De Klerk [2009] and Mesthrie [1992]). A discussion of the characteristics of this variety goes beyond the scope of this study but Mesthrie (1992) deals with this in some detail. The status of these students as NSs or NNSs is an aspect which is impossible to determine, however, without conducting personal interviews and one which made identifying L1 and L2 speakers particularly difficult in this context. The L1/L2 language factor is not the main focus of this study, however, and the sample is primarily regarded as representing novice writers of academic English, regardless of language background.

6.2.2 Research Question 1: Size of productive vocabulary of undergraduate students

Having provided some background information on the overall profile of the students by way of descriptive frequency statistics, in this section my attention is on the research questions driving Phase 1. The first of these was broken down into four sub-questions (see §5.2). In order to answer these, I set out to calculate the mean scores of students on each level of the vocabulary test and their total mean score for the whole test, according to the relevant variables, that is, course, gender, age and home language.

Research Question 1.1: What is the size of the productive vocabulary of students at each level of the VLT, within and across two courses?

Table 6.6 below reflects the mean scores (percentages) and percentiles on the VLT according to the two courses which the students represented, Literature and Law, in answer to Research Question 1.1.

Course type		Level1%	Level2%	Level3%	Level4%	Level5%	Total %
Literature	Mean	91.41	79.50	64.21	66.23	40.41	68.18
	Std dev.	19.32	15.85	24.15	19.48	30.45	18.08
	25th perc	82.35	72.22	50	50	11.11	52.87
	50th perc	94.12	83.33	68.75	66.67	33.33	70.11
	75th perc	100	94.44	81.25	83.33	66.67	83.91
Law	Mean	86.20	69.78	52.32	58.91	27.36	58.75
	Std dev.	12.76	20.48	25.83	20.56	24.35	18.54
	25th perc	82.35	55.56	31.25	44.44	5.56	44.83
	50th perc	88.24	72.22	50	61.11	22.22	58.62
	75th perc	94.12	88.89	75	72.22	44.44	73.56
Total	Mean	88.63	74.31	57.86	62.32	33.45	63.15
	Std dev.	16.34	19.05	25.71	20.36	28.09	18.89

Table 6.6: Scores on VLT according to course (RQ1.1)

In order to examine the differences between academic course and vocabulary levels, a one-way analysis of variance (ANOVA) was applied. The results showed significant differences between course and vocabulary levels: these differences were highly significant at Level 2 ($F_{(1, 296)} = 20.545$, $p < 0.0005$), Level 5 ($F_{(1, 296)} = 16.867$, $p < 0.0005$), Level 3 ($F_{(1, 296)} = 16.694$, $p < 0.0005$) and very significant at Levels 4 and 1 ($F_{(1, 296)} = 9.865$, $p = 0.002$; $F_{(1, 296)} = 7.710$, $p = 0.006$ respectively). It is clear from these scores that the Literature group outperformed the Law group at every level of the VLT. These results also reflect that neither group had reached mastery level (85%) at anything other than Level 1, the 2000-word level. The gap between the two groups' mean scores on Level 2 (high-frequency words) and on Levels 3 (5000-word level) and 4 (the University Word List words or academic words) was particularly noteworthy as these are important levels – students need to have mastered high-frequency words and should be approaching mastery of the academic words if they are to cope with the demands of academic reading and writing (see §3.5.2).

Research Question 1.2: Are there significant differences in the size of the productive vocabulary of male and female students?

In order to address Research Question 1.2, vocabulary scores according to gender were calculated. Table 6.7 reflects descriptive results and the results of the ANOVA for vocabulary differences according to gender:

Gender		Level1	Level2	Level3	UWL	Level5	Total
Male	Mean	84.70	66.58	48.42	56.78	24.53	56.05
N = 123	Std dev	13.551	20.628	24.711	20.088	23.612	18.217
Female	Mean	91.39	79.75	64.50	66.22	39.71	68.14
N = 175	Std dev	17.554	15.815	24.354	19.686	29.319	17.777
F value df 1, 296							
Sig		.000	.000	.000	.002	.000	.000

Table 6.7: Scores on VLT according to gender (RQ1.2)

These results indicate that female students outperformed males consistently across all levels of the VLT, in some cases by more than 10%. ANOVAs revealed that these differences were very significant ($p \leq 0.005$) at all levels. When means according to course and gender were compared, the results were as follows:

Course type	Gender	Mean	N	*Sd*
Literature	Male	64.12	28	19.326
	Female	69.20	111	17.694
Law	Male	53.67	95	17.272
	Female	66.29	64	17.909

Table 6.8: Total vocabulary scores according to gender within courses

It is clear from the table above that gender differences have emerged, particularly in the case of the Law students, where differences between male and female students appear to be greater than in the case of the Literature students. In order to test for significant differences, an independent t-test was included between male and female total vocabulary results within each course. These tests indicated that there were no significant differences between means in the case of the Literature students; however, the difference between male and female scores in the Law course was significant ($F = .243$, $p = 0.000$). Thus these results suggest that reading literature does help male students to build their vocabulary when compared with their counterparts in the Law course. This is an area that could be further researched.

Research Question 1.3: Are there significant differences in the size of the productive vocabulary of students from different age groups?

In answer to Research Question 1.3, Table 6.9 reflects the vocabulary means according to age bands and the results of ANOVAs.

Age Bands		Level1%	Level2%	Level3%	UWL%	Level5%	Total %
≤ 20–29	Mean	86.83	73.27	53.71	57.97	29.15	60.03
(N=138)	Std dev.	13.458	20.597	26.652	20.656	27.124	19.430
30–39	Mean	87.14	71.38	57.63	62.53	30.88	61.72
(N=86)	Std dev.	11.851	18.485	24.823	18.518	25.260	17.563
40–49	Mean	95.01	75.59	62.69	66.67	37.88	67.36
(N=33)	Std dev.	34.304	16.950	23.670	20.926	27.585	18.112
≥ 50	Mean	93.03	83.54	71.99	74.28	57.82	76.03
(N=27)	Std dev.	6.734	15.522	23.345	20.099	31.812	17.792
F value		2.746	3.310	3.279	3.812	5.420	4.403
df 5, 278							
Sig		.019	.006	.007	.002	.000	.001

Table 6.9: Scores on VLT according to age bands (RQ1.3)

All age groups achieved mastery at Level 1. There was a gradual and fairly consistent increase in scores from the youngest age group through to the oldest, with the 50+ group outperforming all the others and approaching mastery at Level 2 (3000-word level). This trend could be partly the effect of experience as well as maturation – the youngest group had simply not had as much exposure to linguistic input as the older groups. Whatever the reasons for these differences in scores, the ANOVAs showed very significant differences between scores of the oldest age group and the rest of the sample at all test levels and test totals.

Post hoc Scheffé tests[26] revealed that the 50+ age group performed significantly better than several of the age groups: the 30 to 39 year age group at Level 2 ($p = 0.042$) and at Level 5 ($p < 0.0005$); the 20 to 29 year age group at Level 3 ($p = 0.01$), at Level 4 ($p = 0.002$) and at Level 5 ($p < 0.0005$); and the 40 to 49 age group at Level 5 ($p = 0.047$). The 50+ group also performed very significantly better than the 20 to 29 year group ($p = 0.001$) and the 30 to 39 group ($p = 0.008$) on the Total vocabulary score.

Research Question 1.4: Are there significant differences in the size of the productive vocabulary of students from different language backgrounds?

The last question to be addressed in this part of the study investigated the effect of language background on vocabulary scores. As this particular sample had only a few representatives of some indigenous languages, it made more sense for statistical purposes to group these according to their language families.

[26] An ANOVA test reflects whether there is an overall difference between groups, but post hoc tests show which specific groups differed. Because post hoc tests are run to confirm where the differences occurred between groups, they should only be run when there is an overall significant difference in group means (i.e. a significant one-way ANOVA result). https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-4.php There are several types of post hoc tests; the Scheffé is used very commonly.

Two South African languages, Venda and Tsonga, are direct descendents of the Narrow Bantu group. Ndebele, Xhosa, Zulu and Swati are further classified into the Nguni group, and Sepedi, Sesotho and Setswana belong to the Sotho-Tswana group of languages (Lewis, Simons and Fennig 2013).

The last table in this section (Table 6.10) reflects the mean scores on the VLT according to home language, and the results of one-way ANOVAs:

Language groups		**Level1%**	**Level2%**	**Level3%**	**UWL**	**Level5%**	**Total %**
Afrikaans	Mean	97.46	83.78	67.06	65.92	39.34	70.49
n = 37	Std dev.	31.42	13.05	19.91	18.53	25.30	14.79
English	Mean	91.70	80.58	69.43	69.84	48.13	71.77
n = 119	Std dev.	10.68	16.84	23.65	18.51	30.10	17.92
Nguni	Mean	81.48	66.77	45.83	54.84	19.86	53.62
n = 54	Std dev.	13.68	21.49	24.01	19.34	19.53	17.32
Sotho	Mean	86.09	67.52	47.72	55.13	20.09	55.13
n = 52	Std dev.	10.74	16.29	23.44	20.43	20.62	15.70
Tsonga	Mean	79.14	57.07	33.52	47.47	13.64	46.08
Venda	Std dev.	16.29	20.95	24.25	18.65	10.93	16.20
n = 11							
Other	Mean	92.02	81.35	57.14	72.22	28.97	66.26
n = 14	Std dev.	8.82	8.04	16.60	12.89	15.59	9.29
F value		3.701	5.880	7.842	5.596	4.893	8.457
df 5, 281							
Sig		.000	.000	.000	.000	.000	.000

Table 6.10: Scores on VLT according to language groups (RQ1.4)

The general trend reflected in Table 6.10 is that the indigenous language groups did not perform as well as the other three language groups. All groups except the Tsonga/Venda group had achieved, or were close to achieving, mastery at Level 1. However, the differences in mean scores at the lower frequency and UWL levels between the English and Afrikaans groups on the one hand, and the other language groups on the other were more pronounced.

The ANOVA results revealed very significant differences at all levels ($p < 0.001$): the scores of the Afrikaans, English and Other groups are markedly higher than the remaining groups, especially on academic vocabulary (Level 4). Post hoc Scheffé tests on the total vocabulary score revealed very significant differences between Afrikaans and all three indigenous language groups (Nguni $p = 0.001$, Sotho $p = 0.003$, Tsonga/Venda $p = 0.003$) and between English and these three groups (Nguni $p = 0.000$, Sotho $p = 0.000$, Tsonga and Venda $p = 0.000$). Post hoc Scheffé tests on the scores for the UWL showed very significant differences between English and Nguni ($p = 0.000$), English and Sotho ($p = 0.001$) and significant differences

between English and Tsonga/Venda (p = 0.016). Possible explanations for these inequalities in vocabulary knowledge are discussed in more detail in §6.2.4.

6.2.3 Research Question 2: Vocabulary size and academic performance

Once the mean scores and the areas of significant difference had been established according to the variables of course, gender, age and home language, Research Question 2 could be attended to.

Research Question 2: What is the relationship between the size of students' productive vocabulary and their academic performance?

Course type		Level1% 2000	Level2% 3000	Level3% 5000	Level4% (UWL)	Level5% 10,000	Total Vocab%	Exam %
Literature	Mean	91.41	79.50	64.21	66.23	40.41	68.18	52.73
	Std dev.	19.32	15.85	24.15	19.48	30.45	18.08	12.22
	25th perc	82.35	72.22	50	50	11.11	52.87	45
	50th perc	94.12	83.33	68.75	66.67	33.33	70.11	53
	75th perc	100	94.44	81.25	83.33	66.67	83.91	62
Law	Mean	86.20	69.78	52.32	58.91	27.36	58.75	56.38
	Std dev.	12.76	20.48	25.83	20.56	24.35	18.54	13.21
	25th perc	82.35	55.56	31.25	44.44	5.56	44.83	47
	50th perc	88.24	72.22	50	61.11	22.22	58.62	57
	75th perc	94.12	88.89	75	72.22	44.44	73.56	66
Total	Mean	88.63	74.31	57.86	62.32	33.45	63.15	54.68
	Std dev.	16.34	19.05	25.71	20.36	28.09	18.89	12.87

Table 6.11: Scores on VLT and Examination according to course (RQ2)

Table 6.11 reflects students' total vocabulary scores and exam results. Neither of the two groups performed particularly strongly in the examination, with the Law students doing marginally better than those in the Literature group. The better performance by the Law students could be the result of the fact that the Law examination paper was made up of a multiple-choice component, counting 50% towards the final mark, and an essay, while the Literature papers required two answers of extended writing. The sample thus comprised students who had written either one or two essays; I did not separate those who had written one essay from those who had written two, and I analysed whatever the Literature students had written, whether essays or extended answers to shorter questions. Unfortunately, I had no access to the breakdown of marks for the two components of the Law exam paper.

In order to test for a relationship between size of vocabulary and academic performance, as reflected in exam scores, Pearson correlations were performed for the whole group of 298 students. A correlation matrix between the various word levels as well as total vocabulary score and the examination scores is

provided in Table 6.12. The results reveal a robust relationship between overall knowledge of vocabulary (Vocab Total%) and performance in the examination (r = .63, p < 0.000). It is interesting to note that knowledge of words at the 5000-word level and the UWL also showed strong correlations with exam performance (.62 and .60 respectively).

N = 298	**Level1 2 000**	**Level2 3 000**	**Level3 5 000**	**Level4 UWL**	**Level5 10 000**	**Vocab Total**
Level2 3000	.589**					
Level3 5000	.490**	.804**				
Level4 UWL	.427**	.766**	.802**			
Level5 10 000	.396**	.712**	.810**	.756**		
Exam	.342**	.570**	.623**	.605**	.551**	.637**

** Correlation is significant at the 0.01 level (2-tailed)

Table 6.12: Correlations between VLT and examination scores

In order to determine which of these word levels best predicted examination scores, a multiple regression analysis was then performed on the group as a whole. The relative contribution of each predictor variable, in this case, the different vocabulary levels, can be assessed in several different ways. 'In the "simultaneous" method (which SPSS calls the Enter method), the researcher specifies the set of predictor variables that make up the model. The success of this model in predicting the criterion variable is then assessed' (Brace et al. 2003:214). Using this method, a significant model emerged: adjusted R^2 = .413 ($F_{5, 292}$ = 42.87, p <0.0005). The predictor variables for academic performance were Level 3 (β[27] = .276, p = 0.046) and the UWL (β = .219, p = 0.052).

When a multiple regression was done separately on each of the courses, using the stepwise[28] method, a significant model emerged in each case: for Literature, the adjusted R^2 = .514 ($F_{5, 133}$ = 135.341 p < 0.0005) and for Law the adjusted R^2 = .498 ($F_{5, 153}$ = 150.121 p < 0.0005). For each group, the Total Vocabulary score was a predictor for exam performance, but the vocabulary levels that were predictor variables for the two courses differed, with Level 3 the strongest predictor variable for Literature and Level 4 (UWL) the strongest predictor variable for Law. This is reflected in Table 6.13:

[27] 'The beta value is a measure of how strongly each predictor variable influences the criterion variable. The beta is measured in units of standard deviation […]. The higher the beta value the greater the impact of the predictor variable on the criterion variable' (Brace et al. 2003:212).

[28] 'Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. Thus, this method should ensure that you end up with the smallest possible set of predictor variables included in your model' (Brace et al. 2003:214)

	Predictor variable	beta value (β)	Sig.
Literature	Level 3	.355	0.01
	Total vocab	.386	0.005
Law	Level 4	.300	0.027
	Total vocab	.426	0.002

Table 6.13: Predictor variables: stepwise method

These results indicate that Levels 3 (the 5000-word level) and 4 (the UWL) are particularly important when it comes to explaining the link between academic performance and vocabulary size; it is these levels which are vital to success in academic study.

6.2.4 Discussion of Phase 1 results

This first phase of the study revealed a demographic picture of the students whose writing made up the Student corpus. There were more women than men in the group and more women were represented in the Literature course than in the Law course. There were more students in total in the Law course (159) than in the Literature course (139). Just under half of the students in the sample were young people under the age of 30, a third were over the age of 30, and the remainder fairly evenly divided between the 40 to 49 year age group and the over 50 group. While a substantial number of students fell into the expected age group for first-year university students, this breakdown reflects the fact that this university caters particularly for individuals who, for whatever reasons, may not wish to attend or may not be in a position to attend a residential university or to continue their studies immediately after completing their schooling.

In answer to Research Question 1.1, there was a significant difference in performance on the VLT in terms of course, with the Literature students outperforming the Law students consistently at every level of the test. Differences between the two groups' mean scores on the 2000-word level and the UWL are particularly worth noting as these are important levels. In addition, research (Cooper 2000; Laufer 1997; Nation and Waring 1997:11; Paquot 2010; Schmitt et al. 2001:56) has revealed that as well as mastering the first 2000 words, students need to have mastered at least the 5000-word level to be able to read fiction and academic texts with ease (see §3.5.1). The results from this study are consistent with previous studies, especially with regard to the Literature students; that is, it is generally accepted that if learners are to read fiction and academic texts with a coverage of 95%, they require mastery of the 2000 high-frequency words of language (core vocabulary) and a beginning knowledge of words at the 5000-word level, as well as some academic words and technical terms (Paquot 2010:9-10). If students do not have this vocabulary knowledge they are likely to find university study particularly challenging. That students were not approaching mastery level at the 5000-word level is a particular cause for concern. One can pick up words through oral discourse (that is, basic interpersonal communicative skills, or BICS) up to the 3000-word level

– but knowledge of words at the 5000-word level is essential if cognitive academic language proficiency or CALP skills are to develop (see §1.2) (Cummins 1999). These are words not typically encountered in everyday conversation but in written texts, and the findings suggest that students who knew these words had had a fairly extensive exposure to written language, that is, through reading, and reading for pleasure beyond the confines of the classroom. The findings also suggest that students who did not know many words at this level would be likely to encounter difficulties in their reading at university.

In answer to Research Question 1.2 and the effect of gender on scores on the VLT, women performed more strongly on all levels of the VLT and in the examination than their male counterparts. This finding may also be related to the fact that Literature students outperformed Law students: there were more women students in the Literature course, and more men in the Law course with a 60:40 ratio of men to women in Law. When means according to course and gender were compared, although gender differences were underlined, particularly among the Law students, no significant differences were found between means in the case of the Literature students; however, the difference between male and female means in the Law course was significant. Thus although differences between male and female students in the Literature course were not significant, these results do suggest that reading literature does help male students to build their vocabulary when compared with their counterparts in the Law course, where differences between genders were significant. It would seem that not only did women students know more words than men in both courses, but also that reading literature does seem to assist male students in building their vocabulary knowledge. This is an area that could be further researched.

In South Africa, where there is not a strong culture of reading, it is possible that gender stereotyping also played a role in these results. There are few reading role models for children in many South African schools, and it may be that reading is not regarded as particularly important, or as a manly pursuit. These results may suggest that the study of literature and the reading of fiction are more attractive to women than to men.

As far as the relationship between age and scores on the VLT and the examination are concerned, and in answer to Research Question 1.3, it is clear that maturity has a positive effect on vocabulary. The two oldest groups achieved mastery (85%) at Levels 1 and 2 (2000-word and 3000-word levels) and had a superior command of academic vocabulary as well as of words at the 10 000-word level; those words which are rare and very often have to do with technical matters and the jargon associated with various professions, are the types of words not typically associated with oral discourse. It may be that exposure to linguistic input accrues over the years with a cumulative effect on these mature students' language proficiency, and especially on their vocabulary. It is also possible that the over-forties, having grown up and attended school before the influence of television and the electronic revolution became all-pervasive,

tended to read more than younger students. The positive effects of extended reading on vocabulary were discussed in Chapter 3.

Conversely, the particularly low vocabulary scores at Level 5 of the 20 to 29-year-olds, and in particular the under-20s, may be accounted for in terms of their immaturity and the fact that they have simply not had time to read as much as the other groups. Whatever the reasons, the ANOVAs showed high significance in the differences between the scores of age bands at all test levels and in the examination.

The questionnaire results revealed that English was the language reported as most spoken as a first language (differing substantially from the national average), an expected finding as this is the language of learning and teaching (LoLT) at the majority of schools in the country and of this university. Two-thirds of the sample consisted of black, Asian or coloured students. With regard to the black students, this may be a reflection of what Coetzee-Van Rooy (2012, 2014) found in her study, namely that black speakers of African languages may report their home language as an African language but English (for these learners usually a non-native variety) as their strongest language, what she explains as a 'functional distinction' between languages in a multilingual repertoire, between those languages used socially and those used at school or university (Coetzee-Van Rooy 2014:133). In Mesthrie's (2010) study, he shows, through a study of sociolinguistic changes among young middle-class South Africans across the ethnic spectrum, how they have, to varying degrees, 'adopted the prestige White middle-class norms, adapting them or resisting change' (2010:3). Investigating a specific aspect of socio-phonetics, Mesthrie (2010:28) found that the group most prone to adopting the norms of white South African English (WSAE) were young black women, with the Coloured group being most resistant to adopting these linguistic traits. Among the Indians, one group resisted 'a full cross-over into the "White" vowel space' (2010:28), while the other group adopted it to some degree. This Mesthrie (2010) believes occurs because ISAE had developed into a sub-variety used by those with 'high levels of English medium education' (Richards and Schmidt 2002:500) long before the end of apartheid, and thus had status within the community.

Black speakers of English, on the other hand, lacked any such reference. Mesthrie (2010:4) argues that this sociolinguistic change reflects a 'degree of deracialisation in South Africa, young peoples' social relalignment and new class formation'. Mesthrie (2010:29) observes that this 'new deracialisation' means that 'labels like "Black South African English"', while still applying to a majority of people are not appropriate for the 'new elites' that he found emerging in his study. He refers to such young black people as 'Black speakers', because their English is a 'cross-over variety with almost no overlap with the norms of their previous generation'.

In the case of the effect of home language on the total vocabulary and examination scores (RQ1.4), the ANOVAs revealed highly significant differences between the language groups and their knowledge of English words. Afrikaans, English and Other groups had greater vocabulary knowledge than the indigenous language groups. In addition, the indigenous language groups had not yet achieved mastery of high-frequency vocabulary. These results suggest that the three indigenous language groups may be the more marginalised groups, an issue which is linked to the inequalities and dysfunction within the school system; such students may not be 'readers' in the sense of individuals who read for pleasure, they may attend schools with fewer written resources, and may have had less exposure to the printed word and to a culture of reading than the other groups. The fact that Afrikaans speakers outperformed the English speakers at Levels 1 and 2, even though the differences were not significant, is surprising at an English university but not easy to explain. As noted in chapter 1 (§1.2), the notion of English as home language and as strongest language has been discussed in some detail by Coetzee-Van Rooy (2011). This may be linked to her claims that while black students in her study specified English as their dominant language, they tended to inflate their own proficiency in English; there were 29 Afrikaans students in the study, and 29 black L1 students. This may thus have affected the mean scores. The distinction Coetzee-Van Rooy (2011) discusses could be fruitfully applied in further research which may reveal interesting insights into this superior performance by Afrikaans speakers at these levels. From Level 3 onwards the pattern reverts to what one would expect, with the English speakers outperforming the Afrikaans speakers as well as those of the other groups.

The indigenous language groups also scored lower in the examination than the other language groups. This may be the effect of the 'two-tier' or bimodal system of education in this country (referred to earlier in more detail in Chapter 1) where the majority of children are still schooled at institutions, rural and in townships, which are poorly managed, lack resources, and have inadequately trained teachers. Such schools may be characterised by low morale and high absenteeism among staff (NEEDU 2013). By contrast, children attending suburban, materially well-resourced and more efficiently administered schools, both government and private, with motivated and well-trained teachers, are at an advantage. Children in the latter system, irrespective of their language background, perform better on the national school-leaving examination and are better prepared, both intellectually and socially, for tertiary education (Fleisch 2008; Howie 2010; Howie et al. 2012; Machet and Tiemensma 2009; Mthombeni 2010; Pretorius and Ribbens 2005:146; Reddy et al. 2012) (see §1.2). This is an issue of some complexity in South African education today, but not one which this thesis addresses directly.

Thus, the findings of the first part of the analysis in Phase 1 answered the first main research question and established variations in the size of a group of South African undergraduates' productive vocabulary.

Research Question 2 investigated the relationship between size of students' productive vocabulary (measured by their scores on the VLT) and their academic performance. Examination scores, for better or for worse, are what allow students access to university in the first place and then to advance up the ladder to graduation. In a sense, in this study the examination was the great leveller – all students, regardless of course or gender, age or language background, fared rather poorly. Somewhat surprisingly, perhaps, the Law group outperformed the Literature group in mean exam performance, with the ANOVA reflecting a significant difference in scores. However, this may be the result of the nature of the examination paper; the Law paper contained a multiple-choice component, which made up 50% of the final mark, and an essay, while the Literature students had to write more extended texts. This may have placed Literature students at a disadvantage, being assessed as they were purely on a more subjective test with no objective component (multiple-choice test) to possibly boost their scores, and also by having to write extended texts under pressure.

Results of Pearson correlations revealed a robust relationship between overall knowledge of vocabulary and performance on the examination. Focusing on students' scores according to course, multiple regressions revealed that Levels 3 (5000-word level) and 4 (UWL) were the predictor variables for Literature and Law students respectively. In other words, these levels were markers of academic performance in these groups: knowledge of words at the 5000-word level predicted success in the exams for the Literature students, while the predictor variable for success among the Law students was knowledge of academic words. The more words one knows at the UWL level, the better one's academic performance is likely to be: in this study, however, not even the 50+-year-olds had approximated mastery at the UWL level. There were significant differences between the two course groups at the academic word level (Level 4) and highly significant differences at the 10 000-word level, which suggests that the Literature students would be more likely to outperform the Law students in academic reading. These levels are thus significant in explaining the link between vocabulary knowledge and academic performance, as measured by the examination scores.

The findings of Phase 1 support research by scholars such as Laufer (1997), Nation and Waring (1997), Reynolds (2005), Coady et al. (1993) and Cooper (1999), all of whom have stressed the importance of size of vocabulary to reading comprehension and academic performance. The importance of vocabulary to reading has long been accepted (Hazenberg and Hulstijn 1996; Laufer 1986; Qian 2002; Read 2007); vocabulary size has been found to predict reading success in second language studies (Cooper 1999; Laufer 1992). Unlike most studies that rely on correlations between vocabulary size and academic performance, however, this study used a more powerful statistical tool to determine the effect of vocabulary knowledge on academic performance, namely multiple regression analysis. Although total vocabulary knowledge was a significant predictor of academic performance, the results also revealed a more nuanced vocabulary profile

for the different courses. Academic performance is dependent to a great extent on reading skill, and vocabulary knowledge is an important component of this skill. Cooper's (1999) South African study found a relationship between L2 students' overall knowledge of academic vocabulary and academic performance. The finding in this study that scores on the 5000-word level and academic vocabulary level in the VLT were predictors of academic proficiency suggests that students with higher scores at these levels will be more proficient readers and more likely to succeed academically.

These findings also underline the usefulness of the VLT in identifying levels at which students have difficulties, allowing teachers to focus on areas of vocabulary that need particular attention (Laufer and Nation 1999).

6.3 PHASE 2 CORPUS ANALYSIS: FUNCTIONAL DISTRIBUTION OF SELECTED VERBS (RQ3) AND THEIR DELEXICAL USE IN MWUS (RQ4)

This section deals with Phase 2, and moves away from discrete word knowledge and measures of breadth of vocabulary knowledge to MWUs and an investigation of one aspect of depth of word knowledge. As much of the corpus analysis was qualitative, results and discussion are presented together in this phase. The analysis involved several steps. The first of these – generating word lists, establishing word frequencies and identifying keywords, generating concordances in order to identify occurrences of the verbs in question, and the classification of the uses of these verbs – made up the first section of the phase. As such, this part of the analysis and its findings are addressed in Research Question 3. The findings of the second part of the analysis, that is, the use of MWUs containing delexical verbs, and the identification and the analysis of errors made in these MWUs by students, are addressed in Research Question 4

6.3.1 Research Question 3: Distribution and functions of *HAVE*, *MAKE* and *TAKE*

This section addresses Research Question 3: How does the distribution of the functions of the three selected verbs compare within and across the Expert and Student corpora? In order to do so, it was essential to capture all occurrences of the three target verbs in question – *HAVE*[29], *MAKE* and *TAKE* – and to establish how frequently they occurred in the different corpora. In achieving this, certain preliminary steps had to be taken. The first of these was using the WordList application of WordSmith Tools to generate word lists for the corpora in their entirety, that is, Expert and Student corpora, and also as sub-corpora. Statistics for the Expert corpus and the Student corpus were generated using the entire list, all 25 texts (that is, 21 literature and four law texts) in the case of the former, and all 298 texts in the case of the latter. These lists

[29] To recap, the capitalised forms, *HAVE*, *MAKE* and *TAKE*, indicate the word or lemma, while the words in lower case, such as *have*, *has*, *had*, indicate those specific word-forms of the lemma.

were not lemmatised because each word-form of the verbs in question (what Stubbs [2001:25] calls the inflectional forms of the lemma or lexeme) was analysed separately.

6.3.1.1 Comparison of Expert and Student word lists

This preliminary step frames the bigger picture of this study. In the Expert and Student corpora, of the six high-frequency verbs which are most commonly used delexically, that is, *DO*, *GET*, *GIVE*, *HAVE*, *MAKE* and *TAKE*, *HAVE* occurred most commonly (Expert corpus: frequency 1744; Student corpus: 2416). *MAKE* occurred 485 times in the Expert corpus and 651 times in the Student corpus; and *TAKE* 287 times in the Expert corpus and 481 in the Student corpus. Although frequencies for *DO*, *GET* and *GIVE* were high (Expert corpus: 895, 88 and 395 respectively; Student corpus: 1343, 455 and 356 respectively), I chose to include *MAKE* and *TAKE* instead of *DO*, which was the second most frequent of the group after *HAVE*, because *MAKE* and *TAKE* are more productive of the type of MWU in which I was interested (see §1.3.2). Table 6.14 below reflects the numbers of each lemma in the respective corpora and the percentage of the corpora these figures represent:

	Expert Lit	**%**	**Expert Law**	**%**	**Student Lit**	**%**	**Student Law**	**%**
HAVE	1345	0.93	399	0.83	1626	1.13	790	1.24
MAKE	350	0.24	135	0.28	493	0.34	158	0.24
TAKE	222	0.15	65	0.13	341	0.23	140	0.22

Table 6.14: Frequencies and percentages of HAVE, MAKE and TAKE in corpora

Given that *HAVE* can act as both a function word (an auxiliary) and as a lexical verb, it is not surprising that it was the most frequent lemma of the three that were selected, although at least one word-form of the other two lexemes occurred very frequently, which justifies the reference to these words as 'high-frequency' verbs (see Table 6.15 below). The percentages above show a similar pattern for the two Expert corpora, and for the two Student corpora, with greater percentages of these three words in the latter than in the former corpora. Further evidence of the higher frequency of these verbs in the Student Lit corpus was provided by log-likelihood calculations which revealed that the Student Lit corpus contained very significantly more occurrences of all three verbs than the Expert Lit corpus (*HAVE*: LL = 29.79, $p < 0.0001$; *MAKE*: LL = 25.97, $p < 0.0001$; *TAKE*: LL = 26.27, $p < 0.0001$)[30]. In the case of the Law corpora, *HAVE* and

[30] The higher the G2 (log likelihood) value, the more significant the difference between two frequency scores. For these tables, a G2 of 3.8 or higher is significant at the level of $p < 0.05$ and a G2 of 6.6 or higher is significant at $p < 0.01$.

- 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
- 99th percentile; 1% level; at $p < 0.01$; critical value = 6.63
- 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
- 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

Rayson's log-likelihood calculator: http://ucrel.lancs.ac.uk/llwizard.html

TAKE were also used significantly more in the Student than in the Expert corpus (LL = 43.92, *p < 0.0001* and 10.92, p < *0.0001* respectively), but *MAKE* was used more in the Expert Law corpus, though this was not significant. The frequency of *HAVE* overall in these corpora corresponds with those of Biber et al. (1999:429), who found that *HAVE* was more frequent in academic prose than any other lexical verb.

6.3.1.2 Comparison of word lists according to course

As one focus of this study is on the comparison of writing by experts and students in two academic courses of study and between student writers in these two courses, the following sections deal with the sub-corpora (see §5.8.1.1 and §5.8.1.2). The word lists of these sub-corpora were compared to determine the frequencies of the target words in each corpus, in preparation for classifying all functions of the verbs and comparing their uses in the corpora [RQ3]. The distribution is illustrated in Table 6.15 below. The two Expert corpora are very similar, both in the rank and in the frequency of word-forms. In the Student corpora, on the other hand, there are remarkable differences in the frequency of *have*, *has* and *had* between Student Lit and Student Law, where there appears to be an overuse of these forms in the Law corpus. This is explored further in §6.3.2.

	Expert Lit						**Expert Law**				
Rank	**Word**	**Freq.**	**%**	**Texts**	**%**	**Rank**	**Word**	**Freq.**	**%**	**Texts**	**%**
21	Have	813	0.56	21	100	23	have	248	0.52	4	100
44	Has	374	0.26	19	90	48	has	108	0.23	4	100
88	make	197	0.14	18	86	96	make	65	0.14	4	100
148	had	122	0.08	15	71	168	had	38	0.08	4	100
170	take	107	0.07	18	86	189	made	35	0.07	4	100
260	made	66	0.05	15	71	192	making	34	0.07	4	100
293	makes	59	0.04	17	81	195	take	34	0.07	2	50
506	making	36	0.02	13	62	502	taking	13	0.03	1	25
514	takes	36	0.02	13	62	586	taken	11	0.02	2	50
575	having	31	0.02	9	43	704	makes	9	0.02	2	25
606	taken	30	0.02	11	52	787	having	8	0.02	3	75
670	taking	27	0.02	11	52	1523	takes	4	-	1	25
739	took	24	0.02	9	43	2754	took	2	-	1	25

Student Lit						Student Law					
Rank	**Word**	**Freq.**	**%**	**Texts**	**%**	**Rank**	**Word**	**Freq.**	**%**	**Texts**	**%**
25	has	752	0.53	133	96	17	have	477	0.75	145	91
38	have	449	0.31	117	84	39	has	253	0.40	109	69
64	had	320	0.22	93	67	103	make	79	0.12	59	37
107	made	190	0.13	85	61	118	take	70	0.11	44	28
114	make	181	0.13	87	63	206	had	39	0.06	29	18
143	take	124	0.09	76	55	215	making	37	0.06	25	16
158	makes	114	0.08	69	50	260	having	31	0.05	25	16
168	having	110	0.08	62	45	263	made	31	0.05	26	16
225	taken	85	0.06	60	43	287	taking	29	0.05	26	16
389	takes	48	0.03	40	29	313	taken	26	0.04	12	14
426	taking	44	0.03	36	26	568	takes	14	0.02	12	8
454	took	40	0.03	27	19	625	makes	12	0.02	12	8
495	making	36	0.03	28	20	1335	took	5		5	3

Table 6.15: Occurrence of target words in sub-corpora

6.3.1.3 Generation of keywords

The next step in the preliminary analysis of these corpora was to generate keyword lists for the entire Student corpus, as well as for the sub-corpora. This process indicated which of the three words were significantly more frequent in one corpus relative to another, and so confirmed statistically the impressions of difference that were gleaned from the wordlist frequencies provided earlier. The keywords in the Student Lit and the Student Law corpora, relative to the reference corpora, namely the Expert corpora, are indicated in Table 6.16 below:

Keyword	Freq.	%	RC Freq.	RC%	Keyness	P
Student Lit vs Expert Lit						
has	752	0.53	374	0.26	134.10	7.97685E-16
had	320	0.22	122	0.08	94.26	3.65601E-15
made	190	0.13	66	0.05	64.09	2.93673%-14
having	110	0.08	31	0.02	47.82	3.02788E-13
taken	85	0.59	30	0.02	28.03	1.16489E-07
Student Law vs Expert Law						
has	253	0.40	108	0.23	26.84	2.18071E-07
have	477	0.76	248	0.52	24.31	8.18266E-07

Table 6.16: Keywords in sub-corpora

The words in these keyword lists are sorted according to keyness, in descending order. Column 2 in the table (Freq.) indicates frequency in the source corpus, with the percentage that is represented by this count relative to all the words in the corpus in column 3. Columns 4 (RC Freq.) and 5 (RC%) represent the

frequency in the reference corpus and the percentage of the corpus this represents respectively. Column 6 contains the keyness value, with column 7 containing the *p* value. The high keyness values and *p* values pertain partly because of the very high frequencies of the words in the fairly large corpora.

This table indicates that *has*, *had, having, made* and *taken* were keywords in the Student Lit corpus, meaning that these words occurred significantly more often than would be expected by chance in comparison with the Expert Lit corpus; *have* was not a keyword in this corpus. In other words, this word was not used significantly more in the Student Lit corpus than in the Expert Lit corpus. This complements the word list findings, which showed that *have* was ranked most frequent in both the Expert Lit corpus and the Student Lit corpus. In the Student Law corpus, *have* and *has* were the only target words that were keywords relative to the Expert Law corpus, indicating that these words were used significantly more frequently relative to the Expert Law corpus.

6.3.1.4 Generation and investigation of concordance lines

The results of the statistical processing of the distribution of the three verbs are presented using two methods that provide two different perspectives on the differences between the corpora – particularly when the two views do not appear to support each other. Firstly, pie charts are used to depict the frequency distribution of each function of each verb relative to one another, that is, relative to the total number of occurrences of each verb in each corpus. Secondly, the log-likelihood (LL) statistic is used to provide an indication of the frequency of each function relative to the number of words in each corpus. The pie charts thus provide a concrete representation of the 'narrower' perspective (percentage of occurrence of each function of the verb) and log-likelihood calculations give an indication of the difference in function frequencies relative to the 'broader' perspective of all the words in the corpus. Chi square calculations could also have been done on the narrower perspective, but it was decided that using pie chart representation for this, and LL for the broader perspective, would make for good complementary ways of addressing the vexed question of statistical difference in linguistic analysis. The discussion of the results reflects both perspectives.

For ease of reference, a summary of the distribution of functions of the three verbs in the Expert and Student corpora is provided in Table 6.17 below:

		Expert Lit	%	Expert Law	%	Student Lit	%	Student Law	%
HAVE	**Aux**	837	62.2	216	54.1	783	48.1	361	45.6
	Operator	2	0.1	-	-	-	-	-	
	Semi-modal	108	8.0	37	9.2	137	8.4	54	6.8
	Lexical	254	18.8	103	25.8	475	29.2	298	37.7
	Core Delex	20	1.4	3	0.7	24	1.4	3	0.3
	Pseudo Delex	124	9.2	40	10.0	207	12.7	74	9.36
	Total	**1345**		**399**		**1626**		**790**	
MAKE	**Phrasal**	9	2.5	5	3.7	4	0.8	1	0.6
	Prepositional	12	3.4	2	1.4	3	0.6	1	/0.6
	Phrasal Prep	6	1.7	-	-	3	0.6	1	0.6
	Lexical	172	49.1	65	48.1	361	73.2	111	70.2
	Core Delex	15	4.2	2	1.4	5	1.0	1	0.6
	Pseudo Delex	136	38.8	61	45.1	117	23.7	43	27.2
	Total	**350**		**135**		**493**		**158**	
TAKE	**Phrasal**	14	6.3	2	3.0	39	11.4	11	7.8
	Prepositional	27	12.1	6	9.2	28	8.2	7	5.0
	Phrasal Prep	-	-	-	-	3	0.8	-	-
	Lexical	112	50.4	26	40.0	102	29.9	63	45.0
	Core Delex	6	2.7	6	9.2	13	3.8	4	2.8
	Pseudo Delex	63	28.3	25	38.4	156	45.7	55	39.2
	Total	**222**		**65**		**341**		**140**	

Table 6.17: Distribution of uses of HAVE, MAKE and TAKE in Expert and Student corpora (basis of pie charts)

6.3.2 Analysis of *HAVE*, *MAKE* and *TAKE*

The following sections address Research Question 3: How does the distribution of the functions of the three selected verbs compare within and across the Expert and Student corpora? The results of the classification of the functions of the three verbs and the comparisons of use in the Expert and the Student corpora are presented according to the sub-questions in Research Question 3, starting with the verb *HAVE*.

6.3.2.1 Analysis of HAVE

This section addresses Research Question 3.1: How does the distribution of the functions of the three selected verbs in the Expert Lit corpus compare with the distribution in the Expert Law corpus? As discussed in Chapter 5, the first step in the analysis was the coding of the functions of this verb (see §5.8.3.1). *HAVE* is a complex verb and has several functions. In this study it was categorised according to the following six functions:

HAVE as auxiliary verb

HAVE as operator

HAVE as semi-modal

HAVE as lexical verb

HAVE as core delexical verb

HAVE as pseudo delexical verb

i. Expert Lit vs Expert Law

The Expert Lit and Law corpora were analysed separately. Once all the concordance lines for *HAVE* in the two Expert corpora had been analysed (a total of 1345 lines for *hav**, *has* and *had* in Expert Lit and 399 in Expert Law), the proportion of verb functions was as depicted in Figure 6.1 (keeping in mind that this refers to the proportion of the total occurrences of *HAVE*, not the entire corpus, and that this is also the case for the pie charts for *MAKE* and *TAKE*[31]).

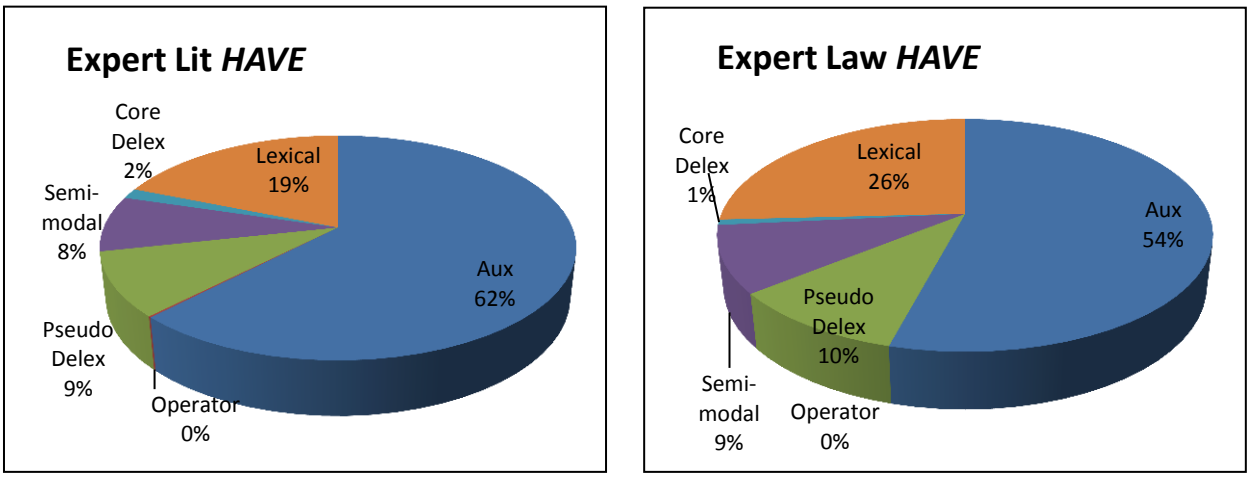


Figure 6.1: Distribution of HAVE functions in Expert corpora

Although the proportional distribution of the verb across the two Expert corpora as illustrated in the pie charts appears to be fairly similar, LL calculations[32] (which compared the frequencies of each verb category relative to the total number of words in each corpus across the two corpora) revealed very significant differences between the use of *HAVE* as auxiliary verb (LL = 11.33, $p < 0.001$) in the Lit corpus relative to the Law corpus. There were very few uses of the verb as core delexical in either corpus, and the two types of delexical MWUs made up only about a tenth of the uses of this verb in these corpora. This is perhaps to be expected in academic writing, although much of the Lit corpus in particular is less formal in tone and is instructional, even conversational, in parts. Biber et al. (1999:1026) found that this sort of 'relatively idiomatic' expression occurred more frequently in conversation, fiction and news reporting than in academic writing, barring the exceptions *take place* (which occurred over 40 times per million words) and

[31] Owing to the rounding up or rounding down of percentages, the figures in charts may not always add up to 100%.
[32] Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis, Lancaster University.
Rayson's Log-likelihood calculator available online at: http://ucrel.lancs.ac.uk/llwizard.html [Accessed 3 January 2014).

have an/no/the effect, *make (any/no) sense, make use of* and *take (the) form (of),* which occurred over 20 times per million words. According to Biber et al. (1999:1028), core delexicals are even rarer, with examples such as *have a look* and *take a look* both occurring fewer than ten times per million words in the Longman Corpus of Spoken and Written English (LSWE) [Biber et al. 1999]). Biber et al. (1999) used. However, it is difficult to ascertain just what the norm is for native speaker writing as not many researchers have quantified this (as noted by Nesselhauf 2005). In Howarth's study, although he found that 'the combined percentages of restricted collocations and idioms are 31 per cent in the LOB subcorpus and 40 per cent in the Leeds (LUSS) corpus' (Howarth 1998b:171), he does not distinguish delexical-type collocations from other types. And although Biber et al. (1999) found that these combinations were rare in academic writing, they do not put a percentage value on this. In the sense that the present study has quantified the occurrences of the various functions of these three verbs, it therefore makes a contribution to this area of linguistic investigation.

Algeo (1995:205) believes that increased use of these expanded predicates in modern English may not only be a result of grammatical changes in English but may also be 'rhetorically or stylistically motivated'. He cites Quirk et al. (1985, cited in Algeo 1995:205), who discuss a number of uses for this type of combination, one of which is to shift the focus of the sentence: the ditransitive expanded predicate (with two objects) focuses on the activity rather than on the participant it affects, such as in *He gave Helen a nudge* versus *He nudged Helen* (Algeo 1995:205). On the other hand, the use of the monotransitive (one object) expanded predicate allows a writer to avoid using a simple unmodified subject verb clause, which often sounds awkward. In illustration of this, Algeo (1995:205) cites Quirk et al.'s (1985:1401) example: *My friend did the cooking* sounds better than *my friend cooked.* Algeo believes that the 'motive for choosing the expanded construction was doubtless always rhetorical' (1995:205). This may well be the case in the Expert sub-corpora in this study, where there were no deviations in the use of these MWUs and where their relative infrequency points to some calculation in their use; whether this is the case in the Student sub-corpora will be investigated in the next section.

A closer look at some concordance lines for *has* used as core delexical verb in MWUs in the Expert Lit corpus suggests a stylistic choice to focus on the activity rather than on the affected participant:

artefact or creation, the film **has a powerful influence** on mass audiences (Exp Lit)
age of this is that the reader **has an inside view** of what the character thinks
f, you will find that the term **has a wide range** of meaning,
king, it is inevitable that it **has an impact** on literature,
larger social context, but also **has an impact** on smaller contexts
aham-Smith (Dr) 2 Harold Bloom **has a very different view** of this speech

As was to be expected, there were no deviations in these combinations in either of the Expert corpora as these were texts from published documents and study materials which had been carefully edited and proofread.

ii. Student Lit vs Student Law

This section addresses Research Question 3.2: How does the distribution of the functions of the three selected verbs in the Student Lit corpus compare with the distribution in the Student Law corpus? The pie charts below in Figure 6.2 reflect the proportion of the total number of occurrences of all word-forms of *HAVE* in the two Student corpora:

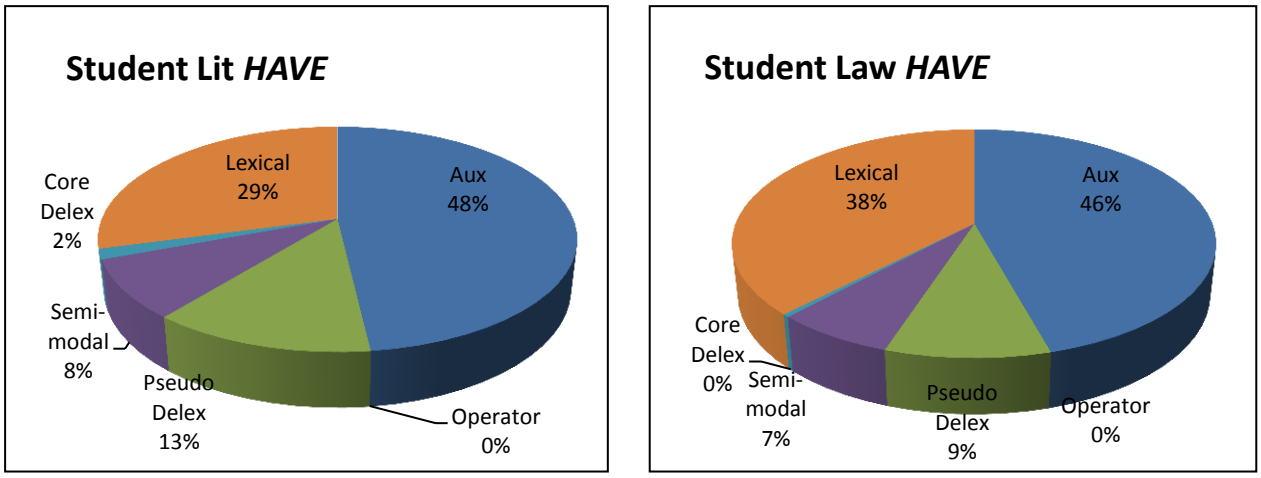


Figure 6.2: Distribution of HAVE functions in Student corpora

As was the case with the Expert corpora, the distribution of the verb categories appears fairly similar proportionally across the two corpora – it does seem, though, that the students in the Law corpus favoured single-word lexical verbs (38% as opposed to 29% in Student Lit corpus), with Rayson's log-likelihood calculator revealing a very significantly greater use in the Law corpus of the lexical verb (LL = 20.88, $p < 0.0001$). On the other hand, Student Lit showed significantly greater use of the core delexical (LL = 5.91, $p < 0.05$).

HAVE as an auxiliary was the most frequent function overall. The second most frequent function in the Student corpora was *HAVE* as lexical verb. All students, but particularly the Law students, tended to use *HAVE* as a lexical verb, mostly in the sense of possessing something, rather than in delexical constructions. In the case of the Student Law, this may have had something to do with the exam topics which required students to write an argumentative essay which in many cases was little more than a list of points for and against a topic, as exemplified in the lines below:

our country as a country which **has money**. Immigrants start d (Stud. Law)

here is other rural areas that **has no electricity** and water.
We elect the one that we know **has a lot of money**, because w
riven by the rich. Governments **have mechanisms** that are desi
ney or poor. These rich people **have networks** everywhere. Cor
r must be forbidden a right to **have tender** from government s

iii. Student vs Expert corpora

Differences between Expert and Student corpora were also explored. This addressed Research Questions 3.3a and Research Question 3.3b.

The Student Lit corpus featured very significantly more uses of *HAVE* as lexical verb (LL = 70.51, $p < 0.0001$) and as a pseudo delexical verb (LL = 21.96, $p < 0.0001$) relative to the Expert Lit corpus. The greater use of the pseudo delexical verbs may have been the effect of informal, even conversational register of some of the writing in the Student Lit corpus; although the Expert Lit corpus employed an informal tone in places, it was essentially didactic and academic in register. This finding is in line with the narrower perspective represented in the pie charts: that is, 9% in the Expert Lit corpus and 13% in the Student Lit corpus represents a substantial difference in frequency.

The pie charts show that the auxiliary made up a considerably larger proportion of the functions of *HAVE* in the Expert Law corpus (54% compared to 46% in the Student Law corpus). However, log-likelihood calculations indicate a significant overuse of the auxiliary in the Student corpus (LL = 7.27, $p < 0.01$). Thus having the two perspectives allows the more complex picture to emerge, showing that although the experts favoured the auxiliary function proportionally more, in terms of the use of this function relative to the overall size of the corpus, the students used it significantly more often. This is largely because Law students used *HAVE* as a whole relatively more often. Students in the Law corpus also used *HAVE* as a lexical verb very significantly more than the writers in the Expert Law corpus (LL = 51.69, $p < 0.0001$), again possibly as a result of the exam questions and the type of arguments students wrote in response.

What the results of these calculations reflect is a tendency on the part of students to overuse *HAVE* both as an auxiliary verb and as a lexical verb when compared to the Expert corpora and, in the case of the lexical verb, in the Student Law compared to the Student Lit corpus as well. In addition, both student corpora produced more examples of the verb in a delexical sense than the equivalent Expert corpus, with a significant difference in the case of the Lit corpora. So while this may suggest a gap in students' vocabulary knowledge, with their generally using *HAVE* to indicate possession, for instance, instead of 'bigger', academic words, the students were also producing or attempting to produce delexical combinations more frequently relative to the Expert corpora. One reason for the greater use of these combinations, particularly in the Student Lit corpus, may lie in the fact that many of them were quoted from the texts provided in the examination, or were current in the content that students had read that semester. Two of

these in the Student Lit corpus, for instance, were *have sex* (33 times) and *have an affair* (17 times), which resulted in a somewhat unrealistic picture of these students' use of this structure. But it may also be that learners are indiscriminate in their use of these MWUs and tend to overuse them, often where a single-word verb would have been more appropriate, as in the examples below, where the more appropriate verb is indicated in square brackets:

assertive and thoughtful person who **has logical thinking** when awkward (Stud Lit) [*thinks logically* …]
initely be why they mostly all **have a dependency** on him to work (Stud Lit)[*depend*]
rate due to the **movement they make** to other places because (Stud Law) [*move*]
good and in deep humility. He **takes conversation** with other (Stud Lit) [*converses*]

In other words, it appears that these combinations are not necessarily 'rhetorically or stylistically motivated' (Algeo 1995:205) in the Student corpora.

6.3.2.2 Analysis of MAKE

As in the case of *HAVE*, the first part of the analysis of *MAKE* was the counting and classification of all occurrences of the verb in the corpora according to the framework below (discussed in detail in §5.8.3.2). There were six categories in all for *MAKE*:

Phrasal verb, e.g. *make up*, *make out*
Prepositional verb, e.g. *make for*, *made of*, *made from*
Phrasal prepositional verb, e.g. *make up for*, be *made up of*
Single-word lexical verb
Core delexical verb
Pseudo delexical verb

These categories differ from those for *HAVE* in several respects: *MAKE* does not act as an auxiliary, as it is a lexical verb only and is used transitively, with rare exceptions such as the colloquial *make out*, in the sense of sexual activity. It may also be a phrasal verb, a prepositional verb or a phrasal prepositional verb.

The concordance lines for each corpus were analysed. The results are discussed in the following sections.

i. Expert Lit vs Expert Law

There were 350 concordance lines for all word-forms of *MAKE* (i.e. *mak** and *made*) in Expert Lit and 135 in Expert Law (see Table 6.17 above). The results are reflected in Figure 6.3 below:

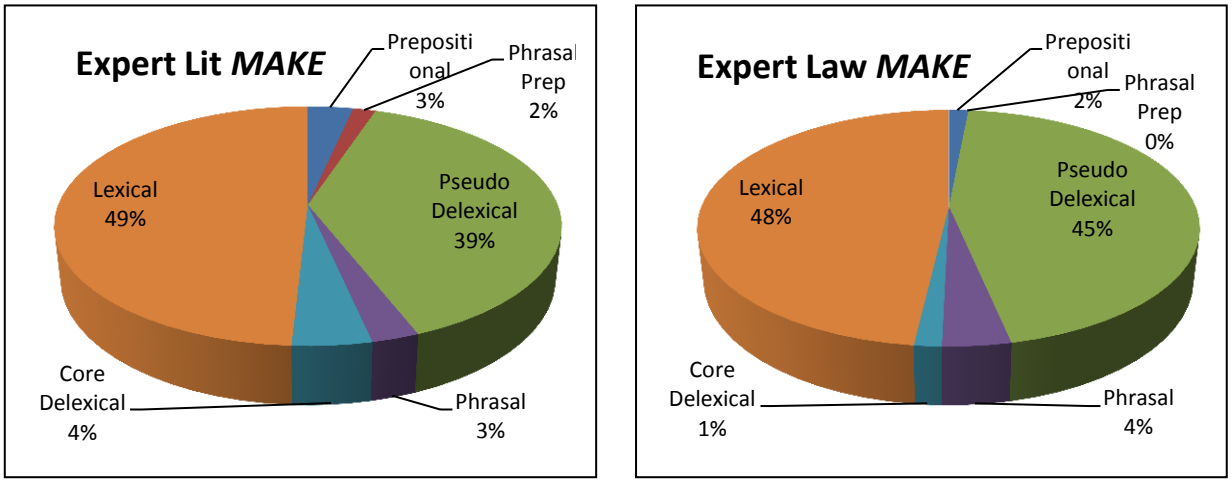


Figure 6.3: Distribution of MAKE functions in Expert corpora

These pie charts suggest that the profile for *MAKE* in the Expert corpora was very similar and log-likelihood calculations revealed no significant differences between these two corpora. There were higher proportions of the verb used in pseudo delexical combinations in both the Expert Lit corpus and the Expert Law corpus than there were for *HAVE,* but lower proportions of core delexicals. This is similar to Algeo's findings (1995:207); he lists more pseudo delexical combinations for *MAKE* than for *HAVE*, although his numbers for core delexical combinations are fairly similar. Biber et al. (1999:1027) also list more examples of these idiomatic or relatively idiomatic expressions (they make no real distinction between core and pseudo delexical uses) for *MAKE* than for *HAVE*, and *MAKE* is noted as being more productive of such combinations (1999:1028).

ii. Student Lit vs Student Law

There were 493 concordance lines for *MAKE* in the Student Lit corpus and 158 in Student Law. The distribution of categories is illustrated in the pie charts in Figure 6.4 below. What these diagrams reflect is that the distribution of categories or proportion of occurrences per category of the verb *MAKE* was almost identical in the two Student corpora.

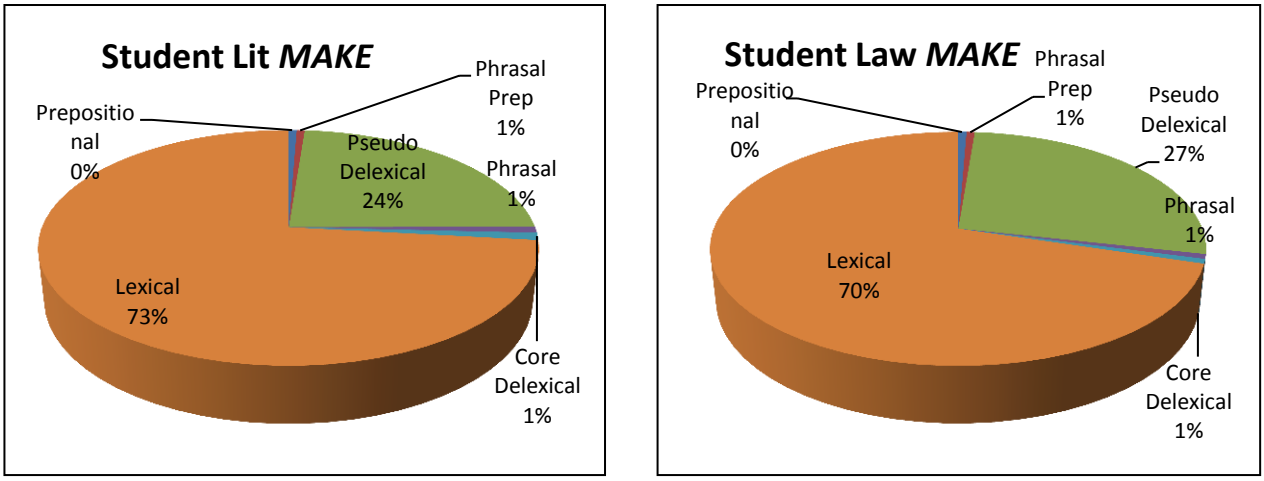


Figure 6.4: Distribution of MAKE functions in Student corpora

As far as the broader perspective of the differences between the Student uses of *MAKE* relative to the corpora as a whole are concerned, log-likelihood calculations revealed significant differences only in the occurrence of *MAKE* as a lexical verb, with this occurring very significantly more in the Lit corpus relative to the Law corpus (LL = 12.38, $p < 0.001$). In both Student corpora, but particularly in the Law corpus, writers tended to use *MAKE* causatively or in the sense of 'force' or 'oblige', particularly in the V+O+inf. combination:

It is times like these that **make me wish** the foreigners w (Stud Law)
reasons and circumstances which **makes me to believe** that corr
words, without naming London, **makes us want** to carry on rea
s able to soften her heart and **make her believe** that he was (Stud Lit)

In the case of Student Lit, this may again have been the result of an overuse of certain quotations from the question paper which featured *MAKE* used as a lexical verb, such as 'good fences make good neighbours', 'make gaps' and 'made himself indispensible'; altogether 33.7% of lexical uses of the verb were made up of some form of these quotations.

iii. Student vs Expert corpora

As far as comparisons between Student and Expert corpora are concerned, addressing Research Questions 3.3a and 3.3b, it is interesting that, given the relatively large proportion of pseudo delexical MWUs in the Expert Corpora (Lit: 38.8% and Law: 45.1%) (see Table 6.17 above for an overview of distributions of all three verbs), this category seems comparatively underrepresented in the Student corpora (23.7% and 27.2% respectively), with most uses of the verb falling into the lexical category. This is the narrow perspective; in exploring the broader perspective, log-likelihood calculations revealed less use of *MAKE* as pseudo delexical in the Student Lit corpus relative to the Expert corpus but this was not significant. Differences in use of the core delexical between Student and Expert Lit were significant, however, with the Student Lit corpus using significantly fewer core delexicals relative to the Expert corpus (LL = 5.12, $p < 0.05$), while the lexical use of *MAKE* in the Student Lit corpus was very significantly greater (LL = 70.59, $p < 0.0001$) than the Expert Lit corpus, with the students using the verb in this way significantly more relative to the Expert corpus. This overuse of the lexical verb in the Student Lit corpus can be explained partly by the very frequent use (over 75 occurrences) of lines quoted from the poem or passage in the exam paper.

The pattern for the use of *MAKE* as a pseudo delexical verb in the Expert Law and Student Law corpora showed a very significant underuse in the Student corpus relative to the Expert corpus (LL = 10.32, $p < 0.01$), and this was also the case with the use of the phrasal verb (LL = 4.17, $p < 0.05$). There were no other significant differences between the Expert and Student corpora.

6.3.2.3 Analysis of TAKE

As in the case of *HAVE* and *MAKE*, the analysis of *TAKE* began by addressing Research Question 3, the counting and classification of all occurrences of the verb in the concordance lines generated for the corpora, according to the framework below. Six categories were identified for *TAKE* (discussed in more detail in §5.8.3.3):

Phrasal verb, e.g. *take up, take on, take off, take over, take apart, take back, take down, take in, take out*

Prepositional verb, e.g. *take NP as, take NP for, take NP from, take to*

Phrasal prepositional verb, e.g. *take away from*

Single-word lexical verb

Core delexical verb

Pseudo delexical verb

i. Expert Lit vs Expert Law

This section addressed Research Question 3.1. In total, there were 222 occurrences of *TAKE* in Expert Lit and 65 in Expert Law. This distribution is illustrated in the pie charts below:

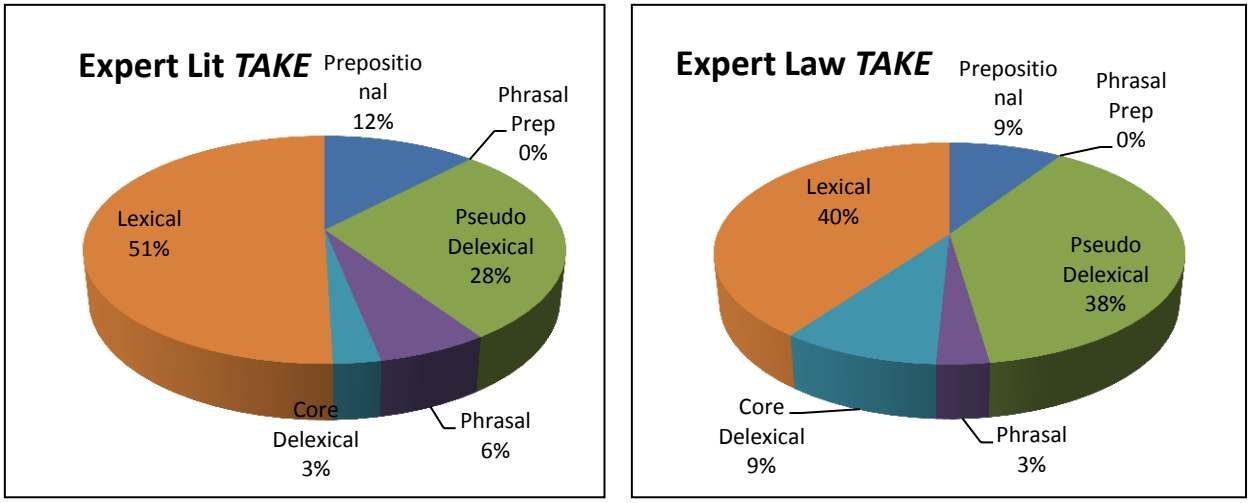


Figure 6.5: Distribution of TAKE functions in Expert corpora

Clearly, this verb was the least frequent of the three, particularly in the Expert Law corpus, where *took* occurred only twice, for instance. Although there were no significant differences in the distribution of functions of the verb in these two corpora, when seen as proportions of the total word count for the corpora (as indicated by the log-likelihood calculations) the Law corpus produced more core delexical and pseudo delexical MWUs than the Lit corpus, while 'narrower perspective' proportions of the verb's occurrences as indicated by the pie charts suggest a higher proportion of lexical verbs for Expert Lit.

ii. Student Lit vs Student Law

As in the case of *HAVE* and *MAKE* above, this section addresses Research Question 3.2. As the least represented of the three verbs, there were 341 occurrences of *TAKE* in the Student Lit corpus and 140 in the Student Law corpus.

As in Expert Law, the word-form *took* occurred only a very few times in the Student Law corpus, five in total. There were more occurrences of this word-form in the Student Lit corpus, a total of 40. Differences in distribution of functions in the two Student corpora were significant only in the use of *TAKE* as lexical verb (LL = 4.05, $p < 0.05$), with a significant underuse in the Lit corpus relative to the Law corpus. This greater use of the lexical verb in the Law corpus can in part be explained by repeated expressions which the nature of the exam questions elicited. Questions required students to provide an argument in support of or refuting a particular proposition: few students were able to work their ideas into logically argued essays, and most simply produced a list of hypothetical examples, many of which were expressed in the form *take for*

example. In addition, a question on corruption resulted in several examples of *taking bribes*, or variations on this theme:

ents are fired or arrested for **taking bribes** of as little as (Stud Law)
they did not work for. Let us **take an example**. A rich man h
conomic, social and political. **Take education** for example, i

The distribution of this verb in the two Student corpora is illustrated in Figure 6.6 below:

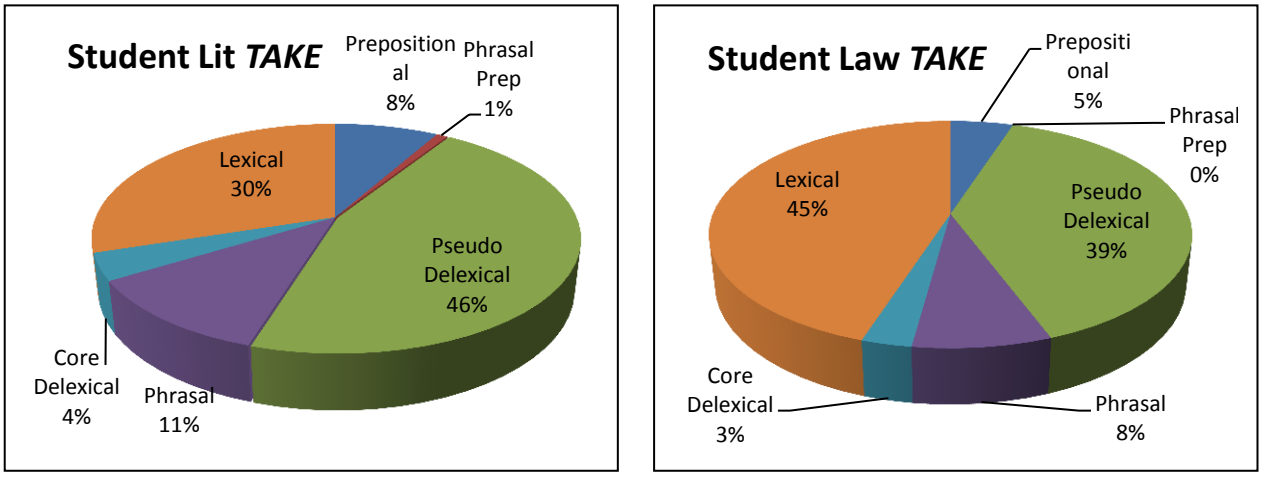


Figure 6.6: Distribution of TAKE functions in Student corpora

iii. Student vs Expert corpora

As far as differences in distributions of categories of verb use in Expert and Student corpora are concerned (RQ3.3a and 3.3b), there were very significant proportional differences in the Student Lit corpus in the use of the verb as pseudo delexical (LL = 41.80, $p < 0.0001$) and as phrasal verb (LL = 12.55, $p < 0.001$). In the case of the overuse of the pseudo delexical by the Lit students, this can be attributed largely to the high use (over 30 occurrences) of the combination *taking advantage*, a quotation from the comprehension passage. Although there was a higher proportion of *TAKE* as lexical verb in the Expert corpus than in the Student corpus, this did not prove to be significant.

There were several significant differences between the distribution of categories of *TAKE* in the Law corpora: differences in use of verb as pseudo delexical (LL = 4.62, $p < 0.05$), as phrasal verb (LL = 4.57, $p < 0.05$) and as lexical verb (LL = 7.15, $p < 0.01$), all reflecting significantly greater use in the Student Law corpus relative to the Expert Law corpus. Although the proportions of pseudo delexicals were very similar (38% and 39% in Expert and Student corpora respectively) in the Law corpora the totals for the use of *TAKE* (Expert Law: 65; Student Law: 140) were relatively much higher in the Student corpus, indicating a much higher use of the verb generally in this corpus (LL = 10.92, $p < 0.001$: it occurred 0.22 times per 1000 words in Student Law, as opposed to 0.14 times per 1000 words in Expert Law). This showed again the value of the two perspectives on the distribution of the verbs; here, distributions that appeared from the narrower

perspective to be fairly similar proved to be significantly different when viewed from the broader perspective, applying the log-likelihood statistic with reference to the total number of words in the corpus. In this case, the overuse of the lexical verb can be put down to the examples mentioned above with regard to the differences between the Student corpora. The overuse of the pseudo delexical verb in the Student Law corpus relative to the Expert Law corpus may be the result of the frequent use of the expressions below which occurred over twenty times in all:

is us who allows corruption to **take place** in our own homes. (Stud. Law)
order to get food and money to **take care** of themselves. They

6.3.3 Research Question 4: Use of MWUs

In this section, the findings of the identification and the analysis of errors made in MWUs produced by students are presented. This section addresses Research Question 4: How does students' use (in terms of frequency and deviance) of selected MWUs compare with the use of these MWUs by expert writers within and across courses?

6.3.3.1 Research Questions 4.1 and 4.2: Frequency of MWUs

This section addresses the first two sub-questions of Research Question 4, that is, Research Question 4.1: How does the use (frequency) of MWUs in the Student Lit corpus compare with their use in the Expert Lit corpus? and Research Question 4.2: How does the use (frequency) of MWUs in the Student Law corpus compare with their use in the Expert Law corpus?

In order to address these questions, all core and pseudo MWUs for each verb were extracted from the concordance lines. These functions of the verbs have been discussed in detail in §6.3.2: there is some overlap in these two sections but I felt it made more sense to discuss all the functions of the verbs together, that is, in §6.3.2. Table 6.18 below (drawn from the relevant parts of Table 6.17) indicates the numbers of delexical MWUs in both Expert and Student corpora:

	Corpus	Core delexical MWUs	Pseudo delexical MWUs	Total	Total no of occurrences of V in corpus	% of V made up by delexical MWUs
HAVE	Expert Lit	20	124	144	1345	10.70
	Student Lit	24	207	231	1626	14.20
	Expert Law	3	40	43	399	10.77
	Student Law	3	74	77	790	9.74
MAKE	Expert Lit	15	136	151	350	43.14
	Student Lit	5	117	122	493	24.74
	Expert Law	2	61	63	135	46.66
	Student Law	1	43	44	158	27.84
TAKE	Expert Lit	6	63	69	222	31.08
	Student Lit	13	156	169	341	49.56
	Expert Law	6	25	31	65	47.69
	Student Law	4	55	59	140	42.14

Table 6.18: Delexical MWUs – all corpora

Once again taking a dual view, and as reflected in Table 6.18 above, the verbs *MAKE* and *TAKE* proved from a narrow perspective, that is, as far as proportions of the occurrence of the three verbs are concerned, to be the most productive of the delexical combinations in this study, with higher proportions of MWUs with *TAKE* in the Student corpora than in the Expert corpora. This is reflected in the last column of the table. From a broader perspective, however, that is, using the log-likelihood calculation based on the occurrence of MWUs with reference to whole corpora, significant differences between the totals for all delexical MWUs (reflected in column 4 of the table) occurred between the Student Lit and the Expert Lit corpora in the case of *HAVE* (LL = 21.34, $p < 0.0001$) and *TAKE* (LL = 44.46, $p < 0.0001$), with very significantly more occurrences in the Student corpus in both cases. As noted above, this difference in the use of *TAKE* as delexical verb was partly the result of the excerpts used in the Literature examination paper: *take advantage (of)*, occurring in the passage for comprehension, featured prominently, making up almost half (48.9%) of all MWUs in the Student Lit corpus. In the Law corpora, the Student corpus featured very significantly fewer occurrences of delexical uses of *MAKE* relative to the Expert corpus (LL = 10.93, $p < 0.0001$

There were no significant differences in the production of these delexical verbs between Expert corpora; in the case of the Student corpora, there were significantly more occurrences of *HAVE* as delexical verb (both core and pseudo) in the Lit corpus relative to the Law corpus (LL = 5.07, $p < 0.05$). These figures indicate that the Literature students produced more delexical MWUs than their Law peers, while the two Expert corpora were more similar in this regard, showing no significant differences in numbers of MWUs produced.

6.3.3.2 Research Question 4.3: Deviant MWUs and errors

In this section, Research Question 4.3 is addressed: How does the use (frequency and errors) of selected MWUs in the Student Lit corpus compare with their use in the Student Law corpus? This was done by investigating the MWUs produced in the two Student corpora. Those which were found to be deviant in some way were analysed for errors which were then categorised according to type (see §5.8.4).

As already noted (§5.8.3.3), a distinction between the terms 'deviation' and 'error' is made in this study. 'Deviation' refers to those MWUs which were in some way problematic, while 'errors' refers to each way in which such MWUs were deviant (see §5 8.4). Of the delexical combinations in the Student corpora, varying proportions were deviant (see Table 6.19 below). As far as errors are concerned, this study investigates the acceptability of errors in collocations and discusses all types of errors, whether grammatical or lexical, but only those which are integral to the MWU itself. In other words, it considers the verb and the noun elements of the collocation and the particles immediately following the noun.

Once all the errors in the deviant MWUs had been coded, they were counted and the percentage of deviant MWUS for each verb was also calculated. Table 6.19 below reflects the number of core delexical MWUs (column CD), the number of pseudo delexical MWUs (column PD), the total number of MWUs, the number of deviant MWUs, the percentage of CD and PD MWUs that were deviant and the number of errors for each of the three verbs in question. At this point, it was clear that the numbers of deviant core delexical MWUs was so small (only six in total) as to make comparisons meaningless; from this point onwards the two categories, core and pseudo, are combined to form one category, delexical MWUs.

	Corpus	CD	PD	Total Delexical MWUs	Deviant MWUs	% Dev MWUs	Errors
HAVE	Stud Lit	24	205	229	33	14.4	51
	Stud Law	3	72	75	15	20.0	16
MAKE	Stud Lit	5	117	122	12	9.8	13
	Stud Law	1	41	42	17	40.4	22
TAKE	Stud Lit	13	156	169	12	7.6	13
	Stud Law	4	55	59	20	45.7	27

Table 6.19: Student corpora – number of deviant delexical MWUs and errors

The percentage of MWUs which were deviant for each verb varied from as low as 9.8% for *MAKE* in the Student Lit corpus, to as high as 45.7% for *TAKE* in the Student Law corpus. In a reflection of the findings in Phase 1, where Lit students consistently outperformed the Law students, the Lit corpus contained a lower percentage of deviant MWUs in general (LL = 14.44, $p < 0.001$), and very significantly fewer errors (LL = 16.39, $p < 0.0001$) relative to the Law corpus. This difference between the corpora was reflected in the

results for both *MAKE* (deviant MWUs: LL = 10.84, $p < 0.001$; errors: LL = 16.64, $p < 0.0001$) and *TAKE* (deviant MWUs: LL = 9.53, $p < 0.01$; errors: LL = 22.71, $p < 0.0001$), with the Lit corpus producing significantly fewer deviant MWUs and errors relative to the Law corpus. Differences between corpora for *HAVE* were not significant except in the number of MWUs in total which these students produced, with the Lit students producing significantly more relative to the Law students (LL = 5.60, $p < 0.05$). As far as the three verbs were concerned, *TAKE* appeared to cause particular difficulty for Law students, with 45.7% of the delexical combinations for this verb in the Law corpus being deviant in some way. These students showed a lack of awareness of the collocational restrictions on both *MAKE* and *TAKE*, revealed in the examples below:

former president of Justice was **making a serious corruption**, (Stud. Law)
he rich people because poverty **takes part**. In most cases of
nes of some Africans also **were taken rescued** by the Spanish

These findings underline what other studies have found – that combinations featuring high-frequency verbs used delexically are notoriously difficult for learners to master (Altenberg and Granger 2001; De Cock and Granger 2004).

6.3.3.3 Categories of deviation

In order to begin to explain these differences in the two Student corpora, one must consider the types of errors made in MWUs in these two corpora. Once the deviant MWUs had been identified, each example was analysed for errors. These were coded according to a framework based loosely on Nesselhauf's '[t]ypes of mistakes in collocations' (2003:232) (explained in §5.8.4). Table 6.20 below provides a breakdown of the errors per category.

	Corpus	ADJ	D	N	P	S	SVC	V	Total errors	Total dev delexical MWUs
HAVE	Stud Lit	1	10	-	10	-	8	22	51	33
	Stud Law	-	4	1	4	-	-	7	16	15
MAKE	Stud Lit	-	4	2	1	-	-	6	13	12
	Stud Law	1	7	1	-	-	1	12	22	17
TAKE	Stud Lit	1	1	2	3	-	1	5	13	12
	Stud Law	1	6	2	2	1	2	13	27	20
TotalDEV %	Stud Lit	2 2.5%	15 19.4%	4 5.1%	14 18.1%	-	9 11.6%	33 42.8%	77	57
	Stud Law	2 3.0%	17 26.1%	4 6.1%	6 9.2%	1 1.5%	3 4.6%	32 49.2%	65	52

Table 6.20: Types of errors

The table reflects the categories into which these errors fell. These are explained in the order in which they appear in the table. Some examples are included in the discussion while the full list of deviant MWUs and errors can be found in Appendix F.

Adjectives (ADJ)

There were relatively very few errors in this category, with only two in each corpus. In the example below, the error lies in the fact that the student has used the adverb *totally* in place of the correct adjective *total,* although it could be argued that although the target adjective (*total*) does fall within the MWU, the actual word used (*totally*) does not:

Babamukuru's approval, she totally **had her dependence**. Even when (Stud Lit)

The error in the second example is somewhat more complex, and was explained as an error of adjectival collocation, in that *living* would collocate more correctly with *good* or *successful* in this context.

with drugs. Drug dealers are **making a wealthy living** out o (Stud Law)

Both these errors suggest a lack of awareness of both the grammatical and the phraseological conventions of the language, and the error of collocation is particularly interesting as it indicates that students may have difficulty with more than simply verb-noun collocations, although confirmation of this would require a specific search for adjective combinations in the corpora.

Determiners (D)

The second category of error concerned the determiner, that is, 'function words which are used to specify the reference of a noun' (Biber et al. 1999:258). Biber et al. (1999) identify three broad groups: predeterminers (e.g. *all, both, half*), central determiners (articles, demonstrative and possessive determiners) and postdeterminers (ordinal numerals and semi-determiners such as *some*, *other*, and cardinal numerals and quantifying determiners). The errors in this study occurred in the second group, the central determiners. Included in this category were deviations concerning articles, pronouns, function words which allow one to refer 'succinctly to the speaker/writer' (Biber et al. 1999:328) and including personal pronouns, possessive determiners, possessive and reflexive pronouns and demonstratives. Speakers of South African indigenous languages find determiners, particularly articles and pronouns, difficult to master and these proved to be problematic for the students in this sample, with 19.4% of errors in the Student Lit corpus and 26.1% in the Student Law corpus falling into this category.

The aspect of determiner use which caused the most trouble for students was the use of the article. The majority of determiner errors fell into this category: 86% in the Lit corpus and 78% in the Law corpus. Examples included cases where the article, in this case *a*, had been omitted:

the rich people because poverty **takes part**. In most cases of (Stud Law)

and examples where the article was present but inappropriate:

neighbours. The neighbours are **having the differences** between (Stud Law)
rich people some of them **are taking an advantage** of poor p

In these examples, the article should have been omitted altogether. Then there were many cases where the wrong article had been used:

tried to escape with her, he **made a huge mistake** of driving (Stud Lit)

The definite article *the* should have been used instead of *a* in this example.

Errors in article use are particularly common in the variety of English spoken by some students in the study, BSAE. Van Rooy (2013:12) confirms that the errors listed here are three possible ways in which article use in this variety differs from the NS varieties: articles are left out altogether (reported on by Gough 1996:61, cited in Van Rooy 2013:12), articles are inserted where they would not be used at all in NS English (Mesthrie 2008, cited in Van Rooy 2013:12), or where articles are muddled, that is, substituted for each other (Greenbaum and Mbali 2002:241-3). In De Klerk's (2006a) analysis of a corpus of Xhosa English she found evidence of a 'loss of distinction between mass and count nouns' (2006a:146) with attendant use of both the definite and indefinite article with non-count nouns. She also found examples of the omission of articles and the insertion of inappropriate articles (De Klerk 2006a:147) and 'gender conflation of pronouns' (2006a:148), all of which occurred in the present study. Minow (2010), focusing on the omission of definite and indefinite articles, showed that the insertion of articles where native varieties would not use an article was the most frequent difference in BSAE, and found that native-like usage increased with proficiency levels.

Other examples in the category of determiner, though less common, involved errors in the use of the personal pronoun:

at all but let justice and law **takes its effect** (Stud Law)

where the plural form of the possessive pronoun *its*, *their,* should have been used.

Differences in the number of errors in the determiner in the two corpora were significant (LL = 6.84, $p < 0.01$), with the Law students making significantly more errors in this category than Literature students. Most errors in the determiner occurred in the use of *HAVE* in the Lit corpus (66.6% of determiner errors) and in the use of *MAKE* in the Law corpus (41.1%). Law students made significantly more errors than Lit students for *MADE* (LL = 5.01, $p < 0.05$) and very significantly more for *TAKE* (LL = 9.12, $p < 0.01$) in this category.

Nouns (N)

Although Nesselhauf (2005:71) found that the noun formed the second most frequently deviant element in her corpus, and in many cases the whole collocation was 'inappropriate' (including stretched verb constructions [SVCs] which should have been single-word verbs), in my study nouns caused relatively fewer problems for the student writers (5.1% of errors in the Student Lit involved the noun, while in the Student Law corpus, this percentage was only slightly higher at 6.1%). All errors in this category concerned the number of the noun, either in cases of verb-noun agreement or article agreement, or where uncountable nouns were used incorrectly in the plural, as in the example below, incorrect because *access* does not have a plural form:

practices. It is the rich who **are having these accesses** in most (Stud Law)

In the example below, *contrasts* should be singular to agree in number with the article *a*:

the beauty of the City. This **makes a contrasts** with its use (Stud Lit)

Differences in the number of errors in this category between the two corpora were not significant in the case of any of the verbs or overall.

Preposition (P)

Prepositions introduce prepositional phrases and act as links between noun phrases and other structures. They may be bound or free, depending on whether or not the choice of preposition is dependent on a specific word in the context. In contrast to the noun, the preposition caused students considerable difficulties in this study, particularly in the case of bound prepositions – after verbs and determiners, errors in prepositions made up the third largest group. There were no significant differences between corpora in the number of preposition errors made with each of the three verbs.

Examples of errors in preposition use included the following:

body else. **The opinion that I have** *on* Uriah Heep is based (Stud Lit) [preposition should be *of*]
Lack of education therefore **has a direct correlation** *to* the (Stud Law) [preposition should be *with*]

Both these errors involve the use of an incorrect preposition in cases where the preposition is bound by the context. These proved to be particularly difficult for students, and are further evidence of a lack of awareness of the idiomaticity of certain combinations in English. In fact, 12 of the total of 20 preposition errors (60%) occurred in cases where the preposition was bound.

Structure (S)

This category of error was taken from Nesselhauf (2003:232), what she defines as 'syntactic structure wrong'. In the event, there was in fact only one MWU in this study which fell into this category:

anything achieve that. If it **take to bribe** somebody in order (Stud Law)

This MWU was explained thus: S – structural error: should read *If a bribe is what it takes?* or *If it takes a bribe (to somebody).* Such an error makes the entire expression difficult to explain in terms of any of the other categories and as such required a more general category which expressed the overall ambiguity of the combination. Despite many difficulties reflected in students' writing, this was the only example of a general structural error in all the deviant MWUs.

Stretched verb constructions (SVC)

SVC is the term Nesselhauf (2005) gives to what I term delexical combinations in this study. She distinguishes these combinations from other verb-noun collocations in that they are formed with light or delexical verbs. In my study, the term refers to delexical combinations which would have been better replaced with a single-word verb. Errors in this category occurred where students used an MWU instead of a more appropriate single-word verb. Most errors in this category occurred in the Lit corpus with the use of *HAVE* (66.6%), but this category made up only 8.4% of the total number of errors in the two corpora. Examples included:

definitely be why they mostly all **have a dependency** on him to work (Stud Lit)

Although this expression with an article appears 21 times in the BNC, in the Lit corpus, a single-word verb such as *depend* would have been more appropriate. In addition, *dependency* is uncountable and as such should not take an article.

this fellow Uriah. He does not **have any trust** *towards* this m (Stud Lit)

In the example above the single verb *trust*, as in *does not trust,* would have been more appropriate.

Some constructions featured an underuse of 'bigger' words, such as *migration* or *immigration*, as in the example below:

rate due to the **movement they make** to other places because (Stud Law)

In this example, the context indicates that a single-word verb such as *migrating* would have been more appropriate, or this expression could have been even more appropriately replaced with the noun *migration*. The effect of such SVC errors is to make students' writing sound rather laboured and unnativelike. Expressions such as this last example also suggest a limited vocabulary.

In Nesselhauf's (2005) study, her findings suggested that her learners did not find these delexical combinations particularly difficult to master; in fact, these combinations seemed to be slightly easier than what she calls non-stretched verb constructions, that is, collocations not formed with delexical verbs. Her findings are supported by Howarth (1998b), who compared collocations with light verbs to other collocations, and found that those with verbs used in the delexical sense did not pose the most problems for students (1998b:181). Nesselhauf (2005:212) believes that these findings suggest that although learners find light/high-frequency verbs difficult they are not the verbs that they get wrong most often; she found that the SVCs with the five most frequent verbs (*have, take, make, do, give*) were not more difficult than SVCs with other verbs. Nesselhauf (2005:113) believes that what makes these SVCs inappropriate appears to be certain 'constraints governing the use of stretched verb constructions' (Nesselhauf 2005:113) of which learners are unaware and which have as yet not been adequately described by researchers. Such constraints in Nesselhauf's corpus included instances where learners misunderstood the import of the noun or verb in the combination, as in the conclusion to an essay, 'To *come to a conclusion* one has to say … that it can't go on like it is now'. In a case like this, *to conclude* would have been more appropriate (Nesselhauf 2005:113).

Verbs (V)

The largest group of errors in both corpora fell into the verb category. In the Lit corpus, 42.8% of all errors were in this category, while the proportion in the Law corpus was 42.9%. Differences in this category between the two corpora were significant for *MAKE* (LL = 9.76, $p < 0.01$), very significant for *TAKE* (LL = 13.02, $p < 0.001$) and, as far as the total number of these errors in the two corpora was concerned, also significant (LL =-9.57, $p < 0.01$). These findings reflect those of Nesselhauf (2003, 2005); she found that the German advanced learners of English in her study made most errors in the verb part of the collocation, although her results, like those in this study, also showed that any of the components of a collocation may be deviant. Nesselhauf (2005) explains her findings by the commonly held belief that verbs, particularly, are the most difficult words for learners to master (Kallkvist 1999, cited in Nesselhauf 2005:77; Kaszubski 2000; Laufer 1991:9; Lennon 1996; Yan 2006).

This is certainly borne out in other research: for instance, Altenberg and Granger (2001), who investigated EFL learners' use of high-frequency verbs, in particular *MAKE*, found that even at an advanced proficiency level, learners revealed that they had great difficulty with these verbs, and when the verb was used delexically this difficulty was compounded (2001:189). And in a study of the use of the verb *DO* in a corpus of Chinese learner English, Yan's (2006) findings confirmed earlier reports (Deng 2003, cited in Yan 2006:40) that Chinese students tended to overuse delexical verbs. In Yan's study, delexical uses of the verb comprised over half the errors students made in their use of *DO*.

Particularly relevant to this study, Nesselhauf (2005:77) found that her learners confused 'high-frequency Germanic verbs, such as *take* and *make*, *get* and *give* etc.', adding weight to this study's premise that these words can be difficult for learners to master.

In the present study, four categories of verb error were identified: collocation, concord, tense and wrong choice of verb. The majority of all verb errors occurred in the collocation category (38.4%), underlining the impression that these writers lacked awareness of collocational restrictions governing these three verbs, as is reflected in the examples from Student Law below:

the rich people because poverty **takes part**. In most cases of (Stud Law)
[*poverty* collocates with *play*, not *take part in]*
business suits who choose to **take part** in corruption, who
[*corruption* collocates with *commit*, not *take part in]*

and from Student Lit:

good and in deep humility. He **takes conversation** with other (Stud Lit)
[*conversation* collocates with *make*, not *take]*
ur with different eyes thus he **makes a conclusion** that will
[*conclusion* collocates with *reaches*, *arrives at*, not *makes*.]

The Law corpus revealed significantly more errors than the Lit corpus in this category (LL = 9.04, $p < 0.01$), and this was an aspect of verb use which the Law students did find particularly difficult, especially in the case of collocations with nouns such as *corruption* (nine errors) and *contribution* (two errors) making up 73.3% of all collocation errors in the Law corpus.

There was only one example of a wrong choice of verb, the use of *hid* in place of *heed* in the example below. This may in fact have been a spelling error; as noted above in §5.8.4, there may be no distinction between long and short vowels such as *heed* and *hid* in the pronunciation of many BSAE speakers:

away, advice that he does not ***take*** **hid** of. Gatsby could also (Stud Lit)

Errors of verb tense and concord made up the bulk of the remainder of verb errors in the two corpora. There were no significant differences in the numbers of errors in these categories between the two corpora. The errors in concord (13 in the Lit corpus and seven in the Law corpus), like the errors in noun and article use, highlight a lack of awareness in these corpora of agreement in general.

In this study I have used 'tense' in the more traditional sense of the term to include aspect. To put it very simply, although tense and aspect both 'relate primarily to time distinctions in the verb phrase' (Biber et al. 1999:460), tense describes the time an action takes place, either in the past or in the present, while aspect denotes whether the activity or state is ongoing or completed. In this study aspect was particularly relevant when dealing with the notion of stative verbs such as *have* that refer to 'unchanging conditions' and are not usually used in the progressive aspect (Richards and Schmidt 2002:34, 513). Tense errors in the verb

predominantly involved examples where the progressive aspect was used incorrectly; the present progressive aspect is correctly used to describe actions which are currently in progress or which are about to take place in the near future (Minow 2010:129). Progressives are a well documented feature of BSAE, as Van Rooy (2013:11) observes: 'The progressive aspect has a long history of scholarly attention in BSAE and many other New Englishes'. Corpus analysis by researchers such as Gough (1996, cited in Van Rooy 2013), Mesthrie (2008, cited in Van Rooy 2013) and De Klerk (2006a:140) has confirmed the 'extension of the progressive to stative verbs' (verbs which describe 'physical situations and mental, attitudinal and perceptual states' (Biber et al. 1999:471) and Van Rooy (2006) shows that 'the progressive is still by far the most frequent with activity verbs' (Van Rooy 2013:11). Statistics in Biber et al. (1999) confirm this finding. In her corpus, Minow (2010:144) found that in her Xhosa data 'the frequency of the progressive decreases with increasing proficiency'.

Van Rooy (2013:11) does note, however, that although findings such as these may suggest that the extension of the progressive 'is a learner phenomenon' which will disappear as proficiency increases, his data (Van Rooy 2006) revealed that the 'underlying semantics of the construction is consistently different from the native speaker prototype of a dynamic event with a limited duration'. He found instead that most uses in his data reflected extended duration: in such cases the 'construction is equally compatible with dynamic and stative predicates' (2013:11). Drawing on his ongoing research into the semantics of the progressive, Van Rooy believes that 'corpus linguistic research on the progressive refutes the simplistic view that the progressive is merely extended to stative contexts' (2013:11). Rather, he argues that it should be regarded as a language variety: 'Any English that has some characteristic features and a population of speakers will be deemed a variety' (Van Rooy 2011:190). He notes that '[t]he distinction between error and conventionalized innovation is one of the crucial issues' facing researchers of New Varieties, and an aspect which sometimes exposes their views to criticism (2011:191). He cites Kachru (1991 in Van Rooy 2011:191), who suggested a distinction 'where deviation refers to a form that habitually differs from those habitually used in Inner Circle/native contexts, but is acceptable in a different (New English) context'.

The progressive aspect is often used as a diagnostic tool in identifying New Varieties. Mesthrie (2008b in Van Rooy 2011:195) notes that '[a] striking and almost universal characteristic among L2 varieties in Africa-Asia is the extension of BE + -ING to stative contexts'. This is certainly the case in descriptions of BSAE where it is regarded as an extension of a new subclass of verbs (De Klerk 2003, cited in Van Rooy 2011:195) and attributed to a tendency among BSAE speakers to disregard the grammatical difference between stative and dynamic verbs in the aspectual system. 'The possibility that the construction has acquired a new meaning, in which the dynamic-stative contrast plays no part, has not been considered previously' (Van Rooy 2011:195–6). In this regard, Van Rooy (2011:196) discusses two issues: are the 'so-called extended progressives' in BSAE simply a 'violation of a constraint on the Standard English construction, which serves

little function' and are they 'regarded as errors or as innovations that are acceptable in high-stake communicative domains?'

In answer to these questions, Van Rooy (2011:196) cites an earlier study (Van Rooy 2006), in which he analysed the progressive construction in BSAE, comparing this to its use by native and foreign (German) language users. His findings revealed that the use of progressives in native and foreign language varieties was similar and reflected textbook Standard English descriptions. In the case of BSAE, however, most of the uses of the progressive emanated from a 'prototype of incompleteness, but extensive duration, for instance:

> People don't attend matches because players **are** not **delivering'** (ICLE-TSN01220) (Van Rooy 2011:196).

In BSAE, unlike in the NS and foreign language uses, the 'temporariness, imminent change and the activity being ongoing or foregrounded at some temporal reference point are not central to the meaning' (Van Rooy 2011:196). The examples from the BSAE corpus 'form a coherent linguistic construction', differing essentially from Standard English (Van Rooy 2011:196). In his study, Van Rooy (2006, cited in Van Rooy 2011:196) found that 17 of the 100 instances of the progressive in ICLE signified a 'so-called permanent state' (Van Rooy 2011:196):

> most of the teenagers and adults **are suffering** from this disease [HIV/AIDS] (ICLE-TSN01116)

Together with perfective or durative meanings (46 examples), the majority of examples sprang from the (other) prototype of BSAE construction. On the other hand, there were only nine examples coming from the SE prototype. This provides clear evidence of 'a new life for the structure BE + ING in BSAE' and he thus 'rejects the description of the extension of the progressive to stative verbs as inaccurate' (Van Rooy 2011:197). Van Rooy's (2006, cited in Van Rooy 2011) study provides evidence that the extension of the progressive to stative verbs should be regarded as an innovation and an aspect of a new variety. He contrasts the findings of a study by Gough (1996, cited in Van Rooy 2011:197), which revealed that the construction was completely rejected by black teachers, with those of a study some years later by Van der Walt and Van Rooy (2002, cited in Van Rooy 2011:197), which found that constructions with 'having' in the progressive aspect were widely accepted. As Van Rooy (2011:197) puts it:

> On both counts, grammatical systematicity and acceptability, therefore, the use of the progressive with different meanings and with a wider range of verbs, can no longer be regarded as a simple error in approximating Standard English norms. Rather, this usage should be regarded as an innovation, already conventionalised in BSAE.

So although the the examples categorised as errors in this corpus of student writing can rightly be regarded as aspects of an acceptable language variety, I did however indicate them as problematic because, although several examples occur with activity verbs, in the particular context the construction can be regarded as

non-standard, and not perhaps what is required in academic writing, and where the simple present or simple past tense would be used instead:

neighbours. The neighbours are **having the differences** between (Stud Lit)
lines sestet where the poet **is having a solution** *of* his problem or hill, were untouched.
He is **having the peaceful feeling** a
ctices. It is the rich who **are having these accesses** in most (Stud Law)
former president of Justice was **making a serious corruption**,

Biber et al. (1999:471–472) observe that verbs such as *feel* which refer to mental or attitudinal states or activities occur frequently (more than ten times per million words) with the progressive. *TAKE* as a verb referring to activities and physical events occurs more than ten times per million words with the progressive. Despite being common with the progressive, however, the use of the verbs in this aspect in these particular examples was clearly incorrect.

As far as errors of concord were concerned, the Law corpus featured more errors relative to the Lit corpus but this was not significant. De Klerk (2006a:143) observes that a 'tendency to simplify concord […] is a frequently remarked-on feature of BSAE' and she investigated this in her corpus of Xhosa English (De Klerk 2006b). The findings in this study certainly support this.

The errors in the verb bear out what has been observed in the marking of assignments and examination scripts, where the overuse of the progressive aspect (present tense) and concord deviations are common problems in verb use among these students. These findings add weight to the argument that these high-frequency verbs demand more attention from course writers and teachers.

To sum up, errors in both Student corpora reflect a lack of collocational awareness in their use of these three verbs. In all cases except preposition and SVC errors, Student Law writers made more errors than the Literature students: differences were significant in the determiner and verb categories, and while both these aspects are known to cause learners difficulties, it is clear that Law students found verb use particularly difficult and showed less awareness of collocational restrictions than the Literature students. Thus the pattern which emerged in Phase 1, where Literature students outperformed Law students in vocabulary, is reflected in the production of MWUs: Law students produced very significantly more deviant MWUs (LL = 13.56, $p < 0.001$) as well as very significantly more errors (LL = 13.94, $p < 0.001$). Although errors in the verb category of both corpora comprised mostly collocation and concord errors, the Law students again made more of these errors, and although not significantly so this again highlights the difference in performance of the two groups.

The errors made by the students in these corpora reflect a paucity of lower frequency vocabulary as well as academic vocabulary, and a lack of depth of knowledge of these kinds of words as well as of high-frequency verbs. An example such as *they mostly all have a dependency on him to work* reflects a poor awareness of the way in which, as a word's function changes, so in many cases does its spelling and form. Dependency is an abstract non-count noun and its use here is ungrammatical. The fragment would have been better phrased as *they mostly depend on him for work*, with the lexical verb *depend* replacing the MWU *have a dependency*. This lack of awareness of inflections and derivations is a characteristic of much of the writing by learners such as these. Granger (1998b:152), for instance, found that learners underused native-like collocations and used instead 'atypical word combinations'. So, although the foreign sounding nature of learner production in her study may have been ascribed to an avoidance of MWUs, Granger in fact found that her learners (advanced learners of English who were native speakers of languages such as Swedish and French) overused them, which made them sound verbose and unnativelike. Granger believes this was probably due to 'an underdeveloped sense of salience and of what constitutes a significant collocation' (Granger 1998b:152).

In a study of Chinese learner English, Yan (2006:40-41) also found that 'Chinese students are always allowing delexical *do* [my emphasis] more freedom to collocate with a wide range of nouns' and that 'learners do not only overuse delexical structures, they also misuse them'. Moreover, in my study the data revealed an overuse of unnativelike collocations such as *make + corruption, have + solution, have + contribution* and *make + conclusion.* Like Farghal and Obiedat (1995:321), who found that their Arabic subjects' 'unawareness of colloquial restrictions of lexical items' led them to produce deviant collocations, this 'unawareness' was clear in many of the errors in my study.

Nesselhauf (2005:72) reports that while some of the collocations in her corpus contained more than one deviation, the bulk of deviant collocations (661) contained only one deviant element or were deviant as a whole, 86 had two deviant elements and 'one even [had] three' (Nesselhauf 2005:72). The corpora in my study revealed a similar profile:

Errors in MWU:	**1**	**2**	**3**	**4**
Literature	42	11	3	1
Law	41	9	2	0
Overall	83	20	5	1

Table 6.21: Errors in corpora

So, as in Nesselhauf's corpus, the majority of deviant MWUs in both corpora contained only one error.

What is clear from these findings is that, in contrast to Nesselhauf and Howarth's studies, students in this study, and in particular the Law students, did have difficulty with delexical combinations. Examples such as *he does not have any trust towards this man* (Stud Lit), *they mostly all have a dependency on him* (Stud Lit) and *the movement they make to other places* (Stud Law) reflect a tendency to use SVCs where a single word would have been more appropriate. This might be partly explained by what Howarth (1998b:186) found; many learners in his study did not understand the 'existence of the central area of the phraseological spectrum between free combinations and idioms. It is in handling restricted collocations that errors of both a lexical and a grammatical nature constantly occur'. Errors stemming from such unawareness were certainly reflected in the Student writing in this study, for instance in the many errors in the collocation of *corruption*. Other deviations include instances where a delexical MWU could express only one of the meanings of the unstretched (single-word verb) verb, such in *one who* ***has no consideration*** *for the* (Stud Law), where *has (no) consideration* does not mean the same as *consider* as although *consideration* is morphologically related to *consider,* the expanded predicate differs semantically in some uses.

Although Nesselhauf (2005) found that her learners did not find SVCs particularly difficult, other researchers have found to the contrary that these are challenging for language learners, particularly those combinations containing light (delexical) verbs (e.g. Altenberg and Granger 2001; Kaszubski 2000; Laufer and Waldman 2011; Wang and Shaw 2008; Yan 2006). Unlike this study, most of these studies have not quantified the difficulty of these combinations nor compared them to other collocations. Howarth (1998b:181) is a possible exception; in his study, the delexical sense of verbs in collocations was the second most common category, though they occurred less in NNS than in NS writing (13% compared to 21% of all the collocations of the patterns studied). He suggested that this was due to avoidance because of students' uncertainty about the appropriate collocability of these words. 'This is not surprising', he says, 'given that even native speakers are capable of confusing delexical verbs (for example, *make an assurance* for *give an assurance*)' (Howarth 1998b:181). What is particularly evident from the analysis of the errors in this study is that while errors occurred in both sub-corpora, there was a marked difference between the two groups of writers, continuing the trend observed in Phase 1

6.3.3.4 Acceptability of deviations

Once all the errors in the deviant MWUs had been identified and classified, they were tested for acceptability by three independent mother-tongue English raters, as explained in Chapter 5 (§5.8.1.5). Depending on the degree to which the deviant MWUs were acceptable or not to native speakers of English, and the extent to which the errors disrupted the meaning and interpretation of the combination (sometimes the original text had to be referred to in order to ascertain from the context the degree of the disruption to meaning), deviant MWUs were categorised as *Marginally Acceptable ?, Largely Unacceptable (*) or Clearly Unacceptable* ***** (these categories are adapted from Nesselhauf 2005).

Table 6.22 below reflects the acceptability of deviant MWUs for each of the three verbs, per corpus. It is clear that the number of deviations in each corpus was very small.

	Corpus	?	(*)	*
HAVE	Student Lit	19	12	2
	Student Law	5	8	2
MAKE	Student Lit	2	8	3
	Student Law	4	6	7
TAKE	Student Lit	3	4	5
	Student Law	3	12	5
TOTAL	Student Lit	24	24	10
	Student Law	12	26	14

? Marginally acceptable; **(*)** Largely unacceptable; *****Clearly unacceptable

Table 6.22: Acceptability of deviations

Again, as in the case of errors, there were significant differences between corpora as far as acceptability of deviant MWUs was concerned. The Law corpus produced very significantly more MWUs which were judged *largely unacceptable* (LL = 9.67, $p < 0.01$) or *clearly unacceptable* (LL = 7.73, $p < 0.01$). This indicates that errors were more likely to obstruct the meaning of MWUs produced by the Law students than in those produced by the Literature students; Law students did produce significantly more deviant MWUs with only one error (LL = 12.43, $p < 0.001$), but this did not affect the overall acceptability of the errors in this corpus.

If a combination was judged *marginally acceptable*, this indicated that although there were clearly problems with the expression, meaning was not totally obscured:

room. He has changed because he **has sympathy** *to* his daughter. (Stud Lit)
got in this world. David **made a lot of mistake** in life
that she and her family would be **taken care** *off* if she does no

Immigration **has had a major effect** *to* the (Stud Law)
with drugs. Drug dealers **are making a wealthy living** out of
lack of corrective action that **take place**. Corrective procedures

Although differences were not significant at this level, fewer of the deviations in the Law corpus fell into the *marginally acceptable* category, suggesting that deviations in this corpus tended to be of a more serious nature; this corpus produced significantly more deviant MWUs in the *largely unacceptable* and *clearly unacceptable* categories. Below are examples of *largely unacceptable* (*) deviations from the two corpora:

assertive and thoughtful person who **has logical thinking** when awkward (Stud Lit)
d everyone has to defer to him **makes all decisions**, and ever
he **has an aggressive behaviour** while he talk

ctices. It is the rich who **are having these accesses** in most (Stud Law)
ideas. *Some they are coming* **to make living**, with some stealing
here in South Africa, he was **take care** of neighbouring countries

The following are examples of *clearly unacceptable* * deviations:

Melanie's father. Lurie says he **have the disgrace** for his who (Stud Lit)
alerting Agnes to stand and **make voice be heard**. The narr
because she has no qualms *for him* **taking the blame** *of* running Myrtle

also been arrests of culprits **have deal** in drug trafficking (Stud Law)
They support the BEE's in **making corruption** for their o
their lives. Nowadays people **take of themselves** and feel s

The difference in acceptability of the MWUs from the two corpora again echoes the results from Phase 1. Not only did the Law students produce significantly more deviant MWUs, but significantly more of these were judged *unacceptable*, either *marginally* or *largely*. The students whose texts made up these two corpora were fairly similar in language background, education and age. Why then should so many more of the delexical verb combinations in the Student Law corpus be judged as unacceptable? The most likely reason is the fact that Phase 1 revealed that Law students had significantly smaller vocabularies than the Literature students.

A closer look at the demographic make-up of the students from the two corpora provides further insights: although the two groups of students were fairly similar in age distribution (see Table 6.2), the Lit corpus was made up predominantly of female students, and the Law corpus had many more male students (see Table 6.1). Women consistently outperformed men on the VLT, indicating that they had larger vocabularies and, in addition to this, the Literature students as a whole performed better than the Law students in the VLT by some 10% (see also Table 6.6). In addition, there were more students who were under the age of 40 in the Law group than in the Lit group (almost 120 were under the age of 40, with only 30 over this age), and older students performed generally better on the VLT. Age and language background factors may also account for students in the Law corpus having a poorer command of English, despite indications that their first language was English in several cases. Of the black students in the Law group, 22 indicated that their home language was English, but ten of these were over the age of 30, suggesting that they may have attended township or rural schools as they would have completed their schooling prior to the opening of state schools to all races.

These are all possible reasons for the more able use of these rather difficult combinations in the texts making up the Student Lit corpus. Although in his study Howarth (1998b:179) found no link between proficiency and the use of these collocations, and regards any deviations as 'highly individual', in the present study the Student Lit corpus was made up of texts from students who may have been intending to major in English at university. This suggests that these were students who read more and who may have fared better in English at school. On the other hand, students in the Law course were required to do less reading and very little writing during the semester, and fewer students in the Law corpus listed English as their home language.

Notwithstanding the differences in numbers of deviant MWUs and errors between the two corpora, the Literature students were also making basic errors in the MWUs they produced. These students wrote extended texts from the beginning of their course and were required to read several full-length texts – novels and plays as well as poetry. The focus of the course is on *how* students write, not simply on *what* they write. Despite this, these students also revealed gaps in their collocational awareness and other errors in the MWUs they produced. This may be partly the result of a lack of practice in writing and limited extended reading in general – although they engaged in more extended writing in their English module than the Law students, relatively speaking they were required to do very little writing during the semester. The fact that there is little teaching of grammar and language in these modules may also exacerbate this situation. A further element of the deviations in the Student corpora is that many errors appear to have become habitual among learners. The examples by both groups of writers provided in this chapter illustrate errors which are common in student writing at this university; article and tense markers are frequently omitted, the progressive aspect is commonly overused and pronouns are often used interchangeably.

Finally, as far as the acceptability of deviations per verb is concerned, more errors in the Lit corpus for *HAVE* fell into the *marginally acceptable* category, while the Law corpus had more *marginally acceptable* errors for *MAKE* and *TAKE* than the Lit corpus. However, there were more errors in the two *unacceptable* categories for both *MAKE* and *TAKE* in the Law corpus, significantly so in the *clearly unacceptable* category for *MAKE* (LL = 6.48, $p < 0.05$) and in the *largely unacceptable* category for *TAKE (*LL = 13.21, $p < 0.001$). Law students had more difficulty than the Literature students with all three verbs, but particularly with *MAKE* and *TAKE*.

6.3.4 Discussion of Phase 2 results

This section has provided a discussion of the findings of the analysis of the corpora, in terms of the distribution of the three verbs in question in the Expert and the Student corpora, and of the production of MWUs by writers in these corpora. The findings of the analysis of student errors in deviant MWUs were also discussed.

As to the first point, the results for *HAVE* showed that, in the case of the Expert Lit and Student Lit corpora, the students significantly overused the verb, both as a lexical and as a pseudo delexical verb. In the Expert and Student Law corpora, students significantly overused the auxiliary verb. The pseudo delexical was also used more by the Law students, although not significantly so. In the case of *MAKE*, the Student corpus revealed differences in the use of the verb as pseudo delexical, used significantly less in the Student corpus, and in the use of the verb as core delexical, again underused by students though not significantly. There was also a significant underuse of the phrasal verb in the student corpus. As far as *TAKE* was concerned, there were significant differences between the Expert Law and Student Law corpora, with students using

the verb significantly less as lexical and pseudo delexical verb, and between Expert Lit and Student Lit, where students used the phrasal verb, the phrasal prepositional and the pseudo delexical verb significantly more than the Expert writers. Differences between the Student corpora were significant only in the overuse of *TAKE* as a lexical verb in the Law corpus.

With regard to deviant MWUs produced by the students and the errors they contained, the results reflected the trend observed in the first phase of the study, in that the Literature students produced more MWUs but significantly fewer deviant ones than the Law students, and these contained significantly fewer errors. Though both corpora featured most errors in the verb category, Law students produced significantly more of these, specifically in the area of collocation, once again indicating that Law students' smaller vocabulary size compared to the Literature students had an effect on their writing. These students in particular made more errors of the type that made their writing sound ungrammatical and unnativelike.

In the following sections, the findings of Phase 3 are discussed.

6.4 PHASE 3 THE RELATIONSHIPS BETWEEN VOCABULARY SIZE, VOCABULARY DEPTH AND ACADEMIC PERFORMANCE

The final phase of the study brought together data from Phases 1 and 2 in order to explore relationships between students' vocabulary size (the breadth of their vocabulary knowledge), their use of the selected delexical MWUs (one very specific measure of their depth of vocabulary knowledge) and their academic performance. In addressing Research Question 5, the relationship between the vocabulary size of the students in the sample selected for this phase (see §5.9) and the number of deviant MWUs they produced and the errors in these were examined. In attempting to address Research Question 6, relationships between this aspect of vocabulary depth and academic performance were explored.

6.4.1 Research Question 5: Relationship between size of productive vocabulary and use of MWUs

The first part of Phase 3 addressed Research Question 5: What is the relationship between the size of students' productive vocabulary and their production of selected MWUs, within and across courses?

As explained in Chapter 5 (§5.9), I used the examination scores as the criterion variable to identify a sample of students whose scores were closest to the mean at each of the three percentile levels (that is, 25th, 50th and 75th percentiles) in the Literature and Law groups. This resulted in a total sample of 60 students, 10 at each level in both groups. Having selected the sample, the texts written by this group of students were extracted and concordance lines for each of the word forms of *HAVE*, *MAKE* and *TAKE* were generated.

These lines were analysed manually and errors were identified in deviant MWUs. Deviant MWUs and errors were counted: Table 6.23 below reflects, together with the mean scores of this group for the 5000-word level, the UWL and the total vocabulary test, the number of MWUs in these concordance lines, the number and the percentage of deviant MWUs and the number of errors:

	Percentile	5000-word level	UWL	Total Vocab%	Total MWUs	Correct MWUs	Deviant MWUs	Deviant MWUs %	Total Errors
		Mean	Mean	Mean					
Lit	25th	56.9	60	68.3	31	25	6	19.0	9
	50th	65.6	68.9	68.7	25	25	0	0	0
	75th	85	78.9	78.9	31	31	0	0	0
Law	25th	44.3	53.9	50.6	13	8	5	28	7
	50th	50	56.7	55.9	8	5	3	37	3
	75th	57.5	68.3	66.6	16	14	2	12	2

Table 6.23: Vocabulary size, MWUs, deviations and errors

Continuing the vocabulary trend observed in the first two phases of this study, the table reflects that the Law students again fared less well than the Lit students. With the exception of the UWL scores at the 25th percentile, the Literature students scored at least 10% higher than their Law counterparts at each percentile level. It is nevertheless clear that both Literature and Law students at the 25th percentile knew fewer words than those at the 50th and 75th percentiles; this knowledge of words increases incrementally from the 25th to 75th percentile in the 5000-word level, the UWL and the Total Vocabulary scores. With regard to the frequency of MWUs there was no discernible incremental pattern, with highest and lowest percentile groups being very similar and the middle groups having lower frequencies, but there was a *qualitative* difference in their production, in that the Lit students at the 50th and 75th percentile levels produced no deviant MWUs.

Taking the results in Table 6.23 at face value, it appears that students who produced only correct MWUs were those Literature students who scored 65% or above at Level 3 (that is those at the 50th and 75th percentiles). Those students who produced deviant MWUs, in both groups, scored below 60% at this level. It may be that knowledge of words at Level 3 is a tipping point for the development of deeper word knowledge, as reflected in a more appropriate use of formulaic language. This would support those scholars who have stressed the importance of this level of word knowledge for reading success at academic level.

In order to test whether there was a relationship between size of vocabulary and the production of deviant MWUs (a measure used as an aspect of depth of vocabulary knowledge), Spearman's rho correlations were conducted between size of vocabulary (5000-word level, UWL and Total vocabulary) and the percentage of deviant MWUs for the sample of 60 students as a group. The results showed modest yet significant

negative correlations between different aspects of vocabulary size and inappropriate use of MWUs, as shown in Table 6.24. These results indicate that the smaller students' vocabulary, the more deviant MWUs they were likely to produce.

Whole Group N = 60	**% Dev MWUs**	**Lit N= 30**	**Law N = 30**
5000-word level	–.27*	–.42*	–.085
UWL	–.29*	–.35*	–.22
Total Vocab%	–.31*	–.40*	–.13

** Correlations significant at .05*

Table 6.24: Correlations: Vocabulary scores and percentage of deviant MWUs

When correlations were conducted for the Lit and Law groups separately, there were significant negative correlations only in the Lit group, between deviant MWUs and vocabulary size at the 5000-word level, with the UWL and with the Total Vocabulary score, as reflected in Table 6.24 above. This reflects the findings of the study by Akbarian (2010), for instance, who also found that students with smaller vocabularies were less likely to develop depth of word knowledge, although it must be kept in mind that in this part of my study both the database and the number of relevant MWUs was very small.

6.4.2 Research Question 6: Relationship between depth of vocabulary knowledge and academic performance

This part of the analysis addressed Research Question 6: What is the relationship between students' production of selected MWUs and their academic performance, within and across courses?

Although exam means at percentiles were used as an organising device for both Research Question 5 and 6, this part of the analysis looks more closely at the exam scores of this smaller sample of students, in relation to their use of delexical MWUs. The higher frequency of deviant MWUs in students' exam answers may negatively influence the coherence of their responses, which in turn may affect the score assigned to them by the examiner. Table 6.25 below reflects mean scores for the exam, the number of MWUs and errors, and the percentage of deviant MWUs:

	Percentile	*N*	*Exam*	MWUs	Dev MWUs	% Dev MWUs	Errors
			Mean				
Lit	25th	10	45.00	31	6	19	9
	50th	10	53.30	25	0	0	0
	75th	10	62.20	31	0	0	0
Law	25th	10	45.90	13	5	38	7
	50th	10	56.80	8	3	37	3
	75th	10	66.00	16	2	12	2

Table 6.25: Phase 3 Sample scores (N = 60)

Although the actual counts are quite low, Literature students at the 25th percentile level produced deviant MWUs at just half the rate of the same percentile in the Law group (19% and 38% respectively). The remaining Literature students in this sample produced no deviant MWUs. Thus there was a notable qualitative difference in the Literature group, in that only the lowest percentile group produced deviant MWUs. This qualitative type of difference was not evident in the Law group: the 50th percentile group fared very similarly to the 25th percentile group, and even among the highest achievers in this course a proportion (12%) of MWUs was deviant. So it seems that, among the students in the Literature group in particular and to a lesser extent in the Law group, students in the higher percentile levels, those who did satisfactorily to well in their exams, used the relevant MWUs more successfully than students who were just passing and those who, with a mark below 50%, were failing their exams.

In order to determine statistically whether there was indeed a relationship between exam scores on the one hand and the number of deviant MWUs and the number of errors on the other, correlations were conducted. Given the small numbers in each group, a non-parametric correlation, Spearman's rho was used. When correlations were conducted for the whole group of 60 students, there were negative correlations between exam scores and deviant MWUs (r_s = −.22, p = 0.090) and between exam scores and errors (r_s = −.22, p = 0.084) but these were not significant. In order to test these results further, correlations were conducted for the two groups separately. The results were much the same: in the Lit group, the correlations were negative but not significant (r_s = −.24, p = 0.185) between exam scores and both deviant MWUs and errors, and this was also the case in the results of the correlations for the Law group (r_s = −.32, p = 0.081and r_s = −.33, p = 0.072 for the correlation between exams and deviant MWUs and errors respectively). As far as type of errors was concerned, most errors occurred in the Verb category in both groups, again supporting the findings from Phase 2, and those of Nesselhauf (2005).

While these correlations may suggest a relationship between deviance of MWUs and errors, and lower exam performance, there were no significant correlations and the numbers in question were very small. As such these findings remain inconclusive, making this an area that merits further investigation.

6.5 CONCLUSION

Addressing each of the six main research questions in turn, this chapter has presented and discussed the findings of the analysis of the data. In Phase 1, the analysis provided interesting findings on the breadth of productive vocabulary knowledge of South African undergraduate students, in answer to Research Question 1. These findings showed that course of study, gender, age and language background were all indicators of vocabulary knowledge, with the Literature students achieving higher scores than the Law students across all levels of the test. Scores on the examination reflected a contrasting picture, with Law students doing somewhat better than their Literature peers, albeit with the whole group of students performing rather poorly. This finding was ascribed partly to differences in the exam structure, where the Law students' marks included an objective test of their ability to read a legal case as well as an essay question to test their writing; the Literature examination contained only questions requiring extended writing. As to gender differences, female students outperformed males at all levels, both on the VLT and in the exam.

As far as age being an indicator of vocabulary knowledge is concerned, the analysis revealed that the group of mature students (above 50 years of age) was the only group to have reached mastery level of the first two levels of the VLT and this group also knew more words at all the other levels. This is a significant finding, given the importance of knowledge of high-frequency vocabulary and a developing knowledge of the 5000-word level to reading comprehension and, by implication, to learning at university. Exposure to language input is a prerequisite for vocabulary development. In order to increase their vocabulary knowledge, young students need to read extensively since exposure to written language increases their chances of learning words at the 5000-word level and beyond. These findings suggest that if students' vocabulary knowledge is inadequate in coping with reading and writing at university level, this could have serious consequences for the throughput, or the number of students who graduate at this university. Prinsloo (2009:18) observes that 'reasons for leaving higher education are usually complex and multiple'; nonetheless, research has shown the importance of teaching and learning factors in success at university.

Finally, that language background was a factor in vocabulary and language proficiency was revealed in the difference in performance between English and Afrikaans speakers on the one hand, and speakers of indigenous languages on the other. The latter group scored relatively poorly in both exam and VLT, results which may well reflect the differences in school background of the language groups in this study.

In answer to Research Question 2 and the relationship between breadth of vocabulary knowledge and academic performance, multiple regression results supported previous research which has emphasised the importance of a knowledge of at least 5000 words to independent reading at university level (Cooper 2000; Laufer 1997; Nation and Waring 1997: Schmitt et al. 2001) by revealing that the 5000-word level, particularly important to reading and understanding fiction, was the strongest predictor variable for academic success for the Literature students. The predictor variable for the Law students, on the other hand, was the UWL. Results showed that some students' knowledge of both these levels was particularly low, suggesting that such students could struggle with university reading and, by extension, writing, unless some intervention is made to increase their vocabulary knowledge. Perhaps the most important finding to emerge from this phase of the study was the suggestion that many students in this sample might not have known enough words to cope with reading academic texts; this could also be a factor in the high failure rates in these two courses.

In Phase 2, in answer to the third and fourth research questions, relating to the distribution and use of the three selected verbs and the number and deviation of the MWUs produced, the analyses of concordance lines found that *MAKE* and *TAKE* produced proportionally more delexical combinations than *HAVE* in all corpora, supporting the findings of studies such as Biber et al. (1999) and Algeo (1995), although students underused *MAKE* in general when compared to the use of this verb in the Expert corpora. There were indications that students had not mastered the collocational uses of *MAKE* and tended to use it more as a lexical verb in its causative sense; in fact, results for all three verbs showed a tendency for students to overuse the lexical function of the verb relative to the Expert corpora. The relative overuse of *TAKE* as pseudo delexical verb by students, particularly those in the Literature group, when compared to the Expert corpora, on the other hand, may have been the effect of quotations from the examination paper.

Deviations in the MWUs produced in the Student corpora supported findings by researchers such as Altenberg and Granger (2001) and Yan (2006) and indicated that delexical combinations do indeed pose particular difficulties for learners, with high percentages of MWUs with *MAKE* and *TAKE* being deviant in the Law corpus, and fewer of these deemed *marginally acceptable* than in the Lit corpus. In this regard, and reflecting findings from Phase 1, Literature students again performed better than Law students, producing a higher percentage of delexical combinations and a lower proportion of errors. Law students made most errors in the categories of verb and determiner, while most errors made by the Literature students occurred in the category of verb. Law students had more difficulties with concord, tense and collocations of the verb, pronouns and articles, which also suggests smaller vocabularies than the Lit students. Verbs presented problems for students in both groups, however; tense and concord errors, as well as determiner errors, are typical of L2 speakers in South Africa, and of speakers of varieties such as BSAE.

These findings reveal that the Law group had more difficulty with delexical MWUs in general, suggesting that more students in this group had a limited awareness of the collocational properties of the three verbs in question. These findings continue the trend which emerged in the first phase of the study, where Literature students scored higher on all levels of the VLT, with the Literature students producing more, and more correct, MWUs relative to the Law students. These findings also provide further evidence of the importance of vocabulary knowledge to academic study. The production of native-like, and correct, delexical MWUs, addressed in Research Question 4, is regarded in this study as a measure of vocabulary depth, and is an aspect of lexical proficiency which is important to academic writing. Errors in the Student corpora (Research Question 4) point to gaps in the vocabulary knowledge of some of the students in this sample and highlight the crucial link between vocabulary breadth and depth; a lack of collocational awareness may have a detrimental effect on their ability to express themselves adequately in academic writing and may also reflect a lack of deep word knowledge.

In the last part of the study, Phase 3, the relationship between the size and depth of students' productive vocabulary knowledge and academic performance, within and across courses, was investigated. Correlations revealed significant negative correlations between size of vocabulary and the production of deviant MWUs for the whole group, suggesting that the smaller the students' vocabulary, the poorer the depth of their vocabulary was likely to be with regard to their use of MWUs; it was significant to note that students who performed at 65% or above on the 5000-word level did not make errors in their use of MWUs. When correlations were conducted for the Literature and Law groups separately, significant, but negative correlations emerged only in the Literature group. The results of these correlations thus support the results for the whole group. These findings corroborate those of Akbarian (2010) (see §3 5.2) who also found that students with smaller vocabularies were less likely to develop depth of word knowledge.

In answer to Research Question 6, when exam scores were correlated with the number of deviant MWUs and the number of errors, both for the group of 60 students as a whole and for the two groups separately, all correlations were negative, albeit not significant. Despite the small sample size and this outcome, the descriptive data suggest that academically weaker students were more likely to produce MWUs that were deviant.

Thus the three phases produced some interesting findings. The trend throughout the study has been for the Literature students to perform more strongly on measures of both breadth and depth of vocabulary knowledge than their counterparts in the Law group. Breadth of vocabulary knowledge was found to have a robust relationship with academic performance, while relationships between size and depth of vocabulary were also found, albeit not as robust. The relationship between depth of vocabulary knowledge and

academic performance was not proved conclusively, but it should be kept in mind that numbers were relatively small.

Chapter 7

Conclusion

7.1 INTRODUCTION

This concluding chapter provides a brief review of what the study set out to do and then moves to a summary of the findings. This is followed by a discussion of the contributions made by the study and the implications of its findings. Recommendations are then made and finally, the limitations of the study are discussed and areas offering scope for further research are identified.

7.2 REVIEW

This thesis reports on a study of the breadth and certain very specific aspects of the depth of students' vocabulary knowledge conducted at a South African open distance learning university. Data were collected from expert writers of English and from undergraduates enrolled in first-level modules offered by the Department of English Studies.

7.2.1 Aims and research questions

The main aims of the study were to measure the size and an aspect of depth of students' vocabulary knowledge and to explore the relationship between the two, and between each of these aspects of their vocabulary knowledge and their academic performance. In order to achieve these aims, the study measured the size of students' vocabulary using a discrete cloze-type vocabulary test and, using corpus linguistic tools of analysis, investigated a very specific aspect of the depth of students' vocabulary knowledge in terms of their production of selected delexical MWUs.

These aims were addressed by attempting to answer six main research questions. Research Question 1 dealt with the breadth of students' vocabulary knowledge, investigating the size of their productive vocabulary and the effects of course, gender, age and language background on vocabulary test scores. Research Question 2 investigated the relationship between the size of students' vocabulary and their academic performance. Addressing these two questions formed the first phase of the study.

In the second phase of the study, the focus shifted to analysis and comparison of the Expert and the Student corpora. Addressing Research Question 3 involved exploring and comparing the distribution of the three high-frequency verbs *HAVE*, *MAKE* and *TAKE* within and across the Expert and Student corpora. Once the distribution of the various functions of these three verbs had been established, MWUs comprising *HAVE*, *MAKE* or *TAKE* used delexically were extracted and analysed in order to address Research Question 4. This included the extraction of deviant MWUs from the Student corpora and the classification of the errors they contained. Differences in distribution and type of error in the two Student corpora, Literature and Law, were investigated and quantified.

The aim of the final phase of the study was to investigate whether there was a link between the two aspects of lexical proficiency, vocabulary breadth and vocabulary depth, and between a specific aspect of depth of vocabulary knowledge and academic performance. In this way, the final phase linked aspects of Phase 1 and 2. For the purposes of this phase of the study, a sub-sample of 60 students' texts was selected using examination scores as the criterion variable. In addressing Research Question 5, the relationship between the size of this group of students' productive vocabulary and the depth of their vocabulary knowledge, measured by their production of delexical MWUs, was investigated. In addressing Research Question 6, the relationship between this aspect of depth of students' vocabulary knowledge and their academic performance was explored.

7.2.2 Main findings

This section highlights the main findings of the study, which were discussed in more detail in Chapter 6. These are presented according to the research questions.

In addressing Research Question 1 in Phase 1, vocabulary scores were considered in the light of four variables. Several findings emerged:

- Perhaps the most significant finding of the analysis in Phase 1 was the relative vocabulary superiority of the Literature students over the Law students. This was a finding that was to form a pattern throughout all three phases of the study. Although the Law students scored higher on the examination than the Literature students, an outcome which may be partly ascribed to the difference in examination format, Literature students scored significantly higher than Law students at all levels of the VLT.
- Mean scores exceeding mastery occurred only at the first level of the vocabulary test, that is, the most common 2000 high-frequency words of English, for both Literature and Law students.
- Female students scored consistently higher than male students across all levels of the VLT and in the examination.

- There was an increase in vocabulary knowledge attendant on age, with the 50+ age group scoring highest at the academic word-level, at 74.28% and, with the exception of scores on the first level of the VLT where the 40 to 49 year age group scored marginally higher, outperforming all age groups across all word levels and the exam.
- As far as language background of the students was concerned, the vocabulary scores of speakers of indigenous languages were significantly lower than those of English and Afrikaans home language speakers across the board. Although this study was not concerned specifically with home language, this difference serves as a reminder that the effects of the institutionalised racial discrimination of apartheid education are still felt by large sectors of the black population in South Africa. Many who are working class or unemployed have no recourse but to send their children to comparatively underresourced and dysfunctional schools; as has been observed in §1.2, in South Africa school background is an influential factor in academic success.

As far as Research Question 2 and the relationship between vocabulary breadth and academic performance were concerned, interesting findings emerged:

- A robust relationship between size of vocabulary and academic performance emerged with a correlation of .63 for the group as a whole. The predictor variable for academic performance for the Literature group was revealed as words at Level 3 (the 5000-word level), while for the Law group the predictor variable was academic vocabulary (the UWL or Level 4). In other words, students' scores at these levels best predicted their chances of academic success. These findings support those of other studies which have shown that both academic vocabulary plus a good knowledge of the first 5000 words of English are vital to successful reading and writing at university level (Laufer 1997; Nation and Waring 1997; Paquot 2010; Schmitt et al. 2001:56).

In Phase 2, the first question which I undertook to answer was whether there was a difference in the distribution of the three verbs between and among the corpora, in terms of both number and function:

- As far as the distribution of these verbs was concerned, there were very few differences between the two Expert corpora, with the only significant difference being in the greater use of *HAVE* as auxiliary in the Expert Lit corpus. The Student corpora produced more occurrences of *HAVE* and *TAKE* as pseudo delexical verbs than the Expert corpora, but fewer in the case of *MAKE*, significantly so in the Student Law corpus.
- Student writers revealed a tendency to overuse the lexical function of all three verbs and *HAVE* as auxiliary.

The second question in Phase 2, Research Question 4, produced the following main findings:

- The Student Lit corpus produced more pseudo delexical uses of all three verbs than the Student Law corpus, significantly more in the case of *MAKE*, as well as significantly more core delexical uses of the verb *HAVE* than the Law corpus. Thus, although students used the verbs more in their lexical function than the Expert writers, they were also producing, or attempting to produce, more delexical MWUs than the Expert writers. This may suggest an unawareness of the formality of academic register and an indiscriminate use by students of these combinations, rather than a conscious decision to use them for stylistic reasons. This is supported by the fact that verb collocation errors and stretched-verb construction (SVC) errors were common in both student corpora, particularly in the Lit corpus. Like Nesselhauf (2003, 2005), I found that the verb was the most problematic element of collocations in students' writing, indicating an unawareness of the collocational properties of the three verbs in question.
- It emerged that, numerically, *TAKE* and *MAKE* were the verbs that were most productive of MWUs in both Expert and Student writing, reflecting what Biber et al. (1999) found in their corpus. Altogether, the Student Lit corpus produced significantly more delexical MWUs than the Student Law corpus, and significantly fewer deviant MWUs with very significantly fewer errors, despite producing more MWUs in general. With regard to the deviant MWUs produced, the verb *TAKE* proved to be the most problematic, particularly for the Law students. In contrast to Nesselhauf's (2005) findings that learners did not find delexical MWUs with high-frequency verbs more difficult to produce than other collocations, the student writers in my study did have difficulty producing these combinations.
- As far as the types of errors were concerned, although the verb proved difficult for both groups of students, bearing out what Nesselhauf (2005) found in her study, Law students found this aspect particularly problematic. These students also made more errors in verb collocation than the Literature students. Once again, the Literature group fared better than the Law group, continuing the trend that had emerged in the first phase of the analysis and strengthening support for the hypothesis that vocabulary breadth has a positive relationship with vocabulary depth, in this case the production of nativelike MWUs.
- Findings on types of errors within deviant MWUs support those of Nesselhauf (2005) and others such as Altenberg and Granger (2001), who have found that the verb is the element which causes most difficulties in these delexical MWUs. Almost half of all errors fell into this category in both Student corpora. Students in my study had particular difficulty with three aspects of verb use, namely collocation, tense and concord. Tense errors were almost entirely cases where the progressive aspect of the present tense was overused, something which is typical of BSAE (Van Rooy 2006, 2013). Concord errors, a typical feature of the English spoken by Afrikaans mother-tongue speakers (Coetzee 2009), were also frequent. Learners also had difficulties with the use of determiners, particularly articles and pronouns. These too are features of English which are known to cause difficulties for speakers of indigenous languages, where the pronoun does not always exist as an

independent word and where there may be no article; errors of both types occurred in the MWUs in this study. These findings support the work of other researchers of South African varieties of English (Coetzee 2009; De Klerk 2006a, b; Van Rooy 2006, 2013) and learner errors (Nel and Muller 2010; Nel and Swanepoel 2010) discussed in this thesis.

- As in Nesselhauf's (2005) study, over half the deviant MWUs in the two Student corpora contained only one error, while eight had three errors and only one had four errors. As far as the acceptability of these deviations is concerned, the Law corpus contained more *largely unacceptable* and *clearly unacceptable* deviations than the Lit corpus. This suggests that the errors in the Law corpus were more serious and global in nature and more inclined to obstruct the meaning and interpretation of the sentence – in fact almost 20% of deviant MWUs in the Law corpus were judged *clearly unacceptable* and over half fell into the two *unacceptable* categories. Fewer MWUs in the Law corpus were judged *marginally acceptable*.
- Both the proportion of delexical MWUs produced and the proportion of deviant MWUs and errors reflected marked differences between writers from the two courses, continuing the trend observed in the first phase of the study; the Lit students produced more delexical MWUs, but fewer of these were deviant and they made fewer errors. This implies a link between knowing more words and being able to use these combinations correctly.

In the third phase of the study, Research Question 5 and Research Question 6 were addressed. Based on a smaller sample of 60 students (30 from Lit and 30 from Law), Question 5 concerns the relationship between breadth and an aspect of depth of vocabulary knowledge, while Question 6 concerns the relationship between depth of vocabulary knowledge and academic performance.

- In answering Question 5, the findings were once again that the Literature students in the sub-samples outperformed the Law students, producing marginally more MWUs relative to the number of occurrences of the three verbs at the 25th and 50th percentiles, though fewer than the Law students at the 75th percentile. None of these differences were significant, however. The difference in numbers of deviant MWUs produced by the two groups was, however, revealed as being significant, with significant negative correlations suggesting a relationship between smaller vocabulary size and the production of deviant MWUs for the group as a whole. There were fewer deviations in MWU use at all three percentile levels in the Lit corpus, with no deviations at the 50th and 75th percentiles.
- In addressing Question 6, although findings revealed no significant relationship between this particular aspect of vocabulary depth, the production of MWUs, and academic performance in this smaller sample, the Lit students used more MWUs and, except for the weaker students at the 25th percentile, used them correctly, while Law students used fewer MWUs and tended to use these incorrectly. This can be linked to the findings for RQ5, that students in the 25th percentiles of both

the Lit and the Law corpora knew fewer words than students in the 50th and 75th percentiles. This suggests an implicit link between academic performance, breadth of vocabulary knowledge and an aspect of depth of vocabulary knowledge, the correct use of MWUs.

7.3 CONTRIBUTIONS OF THE STUDY

The study has made contributions to the fields of both vocabulary studies and corpus linguistics. It has provided findings that are supportive of previous studies, confirming that both breadth and depth of vocabulary knowledge are important to academic writing. In addition, in its findings on the importance of different types of vocabulary to academic success, it has added further credence to the notion that particular academic 'vocabularies' may be relevant to specific course demands, and also to particular socio-cultural contexts, adding support to evidence provided by scholars such as Hyland (2008) and Hyland and Tse (2007). But the study has also added new dimensions to these findings by investigating student writing produced in a complex, multilingual educational context, and in different courses, Literature and Law.

The study has contributed to the current debate on the value of high-frequency vocabulary by underscoring the importance of breadth and depth of knowledge of high-frequency words, words which are often ignored by teachers and even by students themselves (Altenberg and Granger 2001). As has been found in previous studies (De Cock and Granger 2004; Hasselgren 1994; Laufer and Waldman 2011; Lee and Chen 2009; Lennon 1996; Yan 2006), my study has shown that these seemingly innocuous words, when used in formulaic combinations with other words, can cause a multitude of problems for advanced learners at university, regardless of their language background. Although the link between vocabulary and academic performance has been investigated before in a South African university context (Cooper 1999, 2000), this study makes a further contribution by investigating directly, at different levels and across two quite diverse courses, the added dimension of the link between depth of vocabulary knowledge and academic performance, albeit in a very specific and thus narrow manner.

The findings have shown that students have difficulties with regard to one particular aspect of high-frequency word use, that is, the use of delexical verbs in MWUs. In recent years more and more research has investigated these words (e.g. Altenberg 2002; Altenberg and Granger 2001; Hasselgren 1994; Kaszubski 2000; Lee and Chen 2009; Liu and Shaw 2001; Nesselhauf 2004, 2005; Ringbom 1998; Wang and Shaw 2008; Yan 2006) and the findings of this study have added support to evidence that these words may indeed be a '*bête noire'* of learners (De Cock and Granger 2004:233).

This study has added to the body of knowledge by filling a gap in existing research studies on vocabulary and MWUs. Most studies in this area investigate *either* breadth *or* depth of vocabulary knowledge; this

study has considered both and has explored the relationship between them. In this way it has made contributions from a methodological and a theoretical point of view:

- Methodologically, the study has combined the methodologies of discrete-item testing and corpus investigation by using both a breadth of vocabulary measure (a traditional productive cloze-type vocabulary test) and a measure of depth of vocabulary (the production of delexical MWUs in two corpora of academic writing). In so doing, the study has found a relationship between vocabulary size and academic performance and evidence of an indirect link between this aspect of depth of vocabulary knowledge and academic performance.
- Theoretically, the study has used the delexical MWU as a measure of depth of word knowledge, adding to findings from previous studies by Nesselhauf (2005), Howarth (1998b), Altenberg and Granger (2001), Yan (2006) and Wang and Shaw (2008) on these word combinations. It has also reinforced the value of what Biber (2009) refers to as a hybrid approach to corpus investigation, combining as it has both corpus-based and corpus-driven approaches, or what McEnery and Hardie (2012) call corpus-as-method and corpus-as-theory approaches, to good effect.
- In its categorisation and comparison of the distribution of the functions of *HAVE*, *MAKE* and *TAKE* in the Expert and Student corpora, this study has added a further dimension to the understanding of the behaviour of these verbs by revealing the proportions of use of these three verbs by two sets of undergraduate writers. The categorisation and quantification of the errors in deviant MWUs has built on findings by scholars such as Nesselhauf (2003, 2005), finding as in her study that the verb was the most problematic element of the MWUs, and extending her theoretical framework.
- Pedagogically, the study has made a contribution by throwing more light on a very specific group of students' vocabulary knowledge. In its focus on the use of high-frequency words in MWUs by students from a specifically South African background, the study adds to knowledge in this field by providing specific information on the characteristics of the writing required and produced in these courses. It has revealed some of the difficulties such students experience in the use of these verbs. It has also exposed areas of weakness in students' vocabulary knowledge and how this affects their writing and academic performance. In this way, the study has also added information to the body of knowledge on formulaic language by showing how undergraduates in a South African context use certain specific MWUs.

The pedagogical implications of these contributions are discussed below.

7.4 PEDAGOGICAL IMPLICATIONS

Perhaps the most important findings in this study relate to the relationship between breadth and depth of vocabulary knowledge in the context of academic achievement. Evidence of strong relationships between

breadth and depth of vocabulary knowledge, and between size of vocabulary, in particular, and academic performance, have highlighted the importance of word knowledge to academic success. Although both groups of students in this study achieved a mean score of above mastery level of the 2000-word level, it became clear in the corpus investigation that many students had not developed a depth of knowledge of these high-frequency words. This was evident from the type of errors students made in delexical MWUs. Errors of collocation in particular revealed the importance of a deeper knowledge of high-frequency verbs; errors suggested a lack of awareness of how such seemingly simple words behave together with other words, and the restrictions that are frequently imposed by their collocational properties. Thus, in this study, as in those of Lee and Chen (2009), Lee and Swales (2006) and Lennon (1996), learners sometimes appeared to be unaware of the subtleties in the use of these high-frequency words. Errors occurring in MWUs from both corpora suggested a lack of depth of knowledge of words that students did know, combined with a lack of knowledge of lower frequency words; deficits in both breadth and depth of vocabulary knowledge, in other words. A lack of awareness of inflections and derivations also made students' writing come across as unnativelike. In contrast to Altenberg and Granger's (2001) findings, these students tended to overuse delexical combinations particularly where single-word verbs would have been preferable, but they also misused them. A lack of awareness of collocational restrictions was, like that of Farghal and Obiedat's (1995) students, particularly evident from the MWUs produced in the Law corpus. This suggests that knowledge of high-frequency vocabulary, while necessary, is not sufficient, and knowing more words is more likely to lead to greater depth of vocabulary knowledge (Akbarian 2010; Schmitt and Zimmerman 2002; Vermeer 2001).

By investigating differential vocabulary knowledge within and across two different groups of students, the study conducted a more nuanced examination of this knowledge. The findings revealed that the ablest students, those in the top percentile groups in the two courses, achieved mastery of the top two levels of the VLT, and Literature students in this percentile group were close to achieving this at the third and fourth levels, with mean scores of 81.2 and 83.3% respectively. Law students in the 75th percentile lagged slightly behind at these levels with means scores of 75.0% and 72.2% respectively. However, the majority of students in this study clearly need to increase their vocabulary size, and this has implications for their academic success. At this stage of their education, that is, first-year university study, students should ideally have mastered at the very least the first two levels of vocabulary, and have a developing knowledge of Level 3 and academic words (the UWL), providing them with that all-important 5000-word vocabulary which would allow them to read academic texts fairly easily (Cooper 2000; Laufer 1997; Nation and Waring 1997; Schmitt et al. 2001).

Through the comparisons of vocabulary knowledge across the Literature and Law courses, this study also found that different courses may make different vocabulary demands on students: in the case of the

Literature students it was clear that the course demanded the sort of vocabulary knowledge which would allow them to read extended texts with ease, particularly fiction texts, and one which provided them with a base of word knowledge from which they could build up a wider vocabulary, which would then include a fair knowledge of academic words and those words encountered only rarely, the 10 000-word level. Cunningham and Stanovich (2001, 2003) have claimed that reading makes people 'smarter' (2003:34), finding that 'reading volume is a very powerful predictor of vocabulary and knowledge differences' (Cunningham and Stanovich 2001:142). It is possible that the differences in breadth measures of these two groups of students may to some extent have been a reflection of the reading demands made on them, and of their reading ability and of their exposure to reading. Literature students were required to do more extended reading than Law students, and possibly more extensive reading as well. This suggests that the Literature students may have read more than their Law peers, had better reading skills and were as a result more likely to develop other 'aspects of verbal intelligence' (Cunningham and Stanovich 2001:143); or as these scholars express it, they were more likely to 'get richer' while less successful readers 'get poorer'. This points perhaps to a lack of focus on language development, specifically vocabulary development, in courses such as the Law course; somewhat ironically, this course was designed specifically to improve students' reading and writing skills in that discipline.

In this study, academic vocabulary proved to be the predictor of exam success in the case of the Law students. The findings may suggest that the demands of this course of study were somewhat different from the Literature course: there was less focus on reading extended texts, such as textbooks and literary works, and more attention on reading subject-specific material such as law cases, which are often couched in fairly opaque academic words, even though students may have been required to read many of these cases. Errors were judged to be less acceptable in the Law corpus, although writers in both corpora made errors which are typical of South African speakers of other varieties of English such as BSAE, and of speakers of English as a second or third language. These findings suggest that teaching a generic vocabulary list may not be useful in all circumstances: different courses may be better served by being taught specific and different kinds of words (Hyland 2013; Hyland 2008; Hyland and Tse 2007).

It may well be that important reasons for the differences in vocabulary knowledge, both size and depth, between the courses are students' motivation and their history of reading. It is possible that students in the Literature strand may be planning to major in English. This suggests a group of students who may have read more and may be more interested in the language and literature for itself, different perhaps from those in the Law group, where students were obliged to complete a module in English as one of their course requirements. Thus Literature students may have been intrinsically motivated by their own desires to enjoy English for its own sake, while the Law students were more likely to have been extrinsically motivated by the desire to complete their qualification.

7.5 RECOMMENDATIONS

One main issue has emerged from the findings in this study on vocabulary breadth and depth: the need to increase vocabulary size and depth among university students such as those in this study. The challenge is how best to do this.

Signs that some university students have not yet mastered the vocabulary requirements for Grade 6 First Additional Language (FAL) learners in South African schools (CAPS n.d.) (see §3.5.1) is worrying. It appears that, as Lennon (1996) found, learners require more time in the classroom to be spent on vocabulary development; particularly the specific teaching of aspects of so-called 'simple' high-frequency words and their collocational properties, and more practice in the use of these words. The findings on vocabulary size in particular in this study suggest that South African schools should be doing a better job of producing learners who have adequate breadth of vocabulary knowledge, a knowledge that will allow them to develop the depth of word knowledge they need for university study. Although the new curriculum document (CAPS n.d.) certainly places emphasis on vocabulary development, raising as it has the vocabulary requirements for FAL learners at all grade levels in primary school, the vocabulary size of the students in this study underlines the importance of the implementation of requirements such as these if learners are to succeed academically. These CAPS requirements refer to the number of words learners at primary school should know, yet learners at the lowest percentile levels in this study were not showing mastery of even the first 1000 words of English. The problem seems then to lie in the practical classroom implementation of such theoretically well supported curriculum documents. More focus should be paid to vocabulary development in schools; as noted in §1.2, the culture of reading is not well entrenched in many schools or communities and without reading, vocabulary knowledge is unlikely to grow. Reading and vocabulary knowledge, both breadth and depth, feed naturally into good writing skills. These findings have underlined the importance of ensuring that schools provide learners with a sound foundational knowledge of the basic vocabulary of their language of instruction, without which they will find it difficult to develop the depth of their vocabulary knowledge. This can only be achieved if more time in the classroom is spent on literacy activities, starting right from pre-primary and foundation phases.

One way of doing this is to encourage more exposure to texts and reading in the classroom and more extensive reading. The Department of Basic Education (DBE) is very aware of the low literacy levels, evident from both national systemic evaluations as well as the international SAQMEC and PIRLS results. Current annual national assessments (ANAs) in literacy and numeracy have also highlighted this. In 2008, the DBE launched the *Foundations for Learning* initiative in an effort to get reading back on track, and various interventions have been implemented at provincial level. For instance, in Gauteng Province the Gauteng Primary Literacy and Maths Strategy (GPLMS) was implemented in 2011, targeting about 720 of the lowest performing schools in the province in order to improve literacy levels. Both the current and the previous

minister of the DBE have demonstrated their awareness and concern about the problem of low literacy levels. The introduction of the ANAs was one result of this recognition of the need to monitor literacy levels more closely and to encourage schools to raise their literacy levels. In keeping with this awareness, the new CAPS document places a great deal of emphasis on early reading development.

The South African Department of Basic Education also receives support in its drive to raise literacy levels and promote a reading culture from several NGOs such as The Centre for the Book (the National Library of South Africa), Read, Educate, Adjust, Develop (READ), The Molteno Project, the Early Learning Resource Centre (ELRU) and Project Literacy, for instance (Yeh 2004). The National Library also runs several programmes to support the development of children's literature in South Africa, and to 'instill a passion for reading'.[33]

In addition to a focus in teaching on increasing students' vocabulary size, there is a need for the development of a metalinguistic awareness among learners. This becomes a didactic issue and demands attention in the classroom to how one teaches vocabulary as well as what vocabulary one teaches. Lennon (1996:34) suggests an 'inculcation of metalinguistic awareness' to encourage learners to 'focus on language problem areas'.[34] In other words, teachers should try explicitly to raise learners' awareness of their own vocabulary development and thereby help them to increase both the breadth and depth of their vocabulary knowledge. Corpora and concordancing software present new ways of engaging students in such activities, and there are many online resources. Examples of such resources are the Kibbitzer pages developed by Tim Johns and the MICASE Kibbitzer (Reppen 2010:34). These sites provide teachers and advanced language learners with information about words derived from the results of corpus searches. Tim Johns' Kibbitzers (1994, cited in Reppen 2010:34) use an excerpt from student writing to contextualise the particular aspect of language being dealt with. Kibbitzers may provide examples from concordances, or exercises in which answers are available through separate links. Johns also arranged his Kibbitzers in different categories, such as vocabulary, grammar, first language and so on. Kibbitzers include exercises on where to use particular words, for instance, and these could be usefully applied to MWUs.

The MICASE Kibbitzer uses the Michigan Corpus of Academic Spoken English (MICASE) as its corpus. This was inspired by Johns' (1994) Kibbitzers, but the MICASE Kibbitzers differ in format, and are like 'mini research projects' (Reppen 2010:35) and require the teacher to adapt them somewhat for classroom use. Both Kibbitzers could theoretically be very useful in teaching in my own situation; online activites could be

[33] http://www.nlsa.ac.za/index.php/childrens-literature-rogramme (accessed 23 October 2014)

[34] http://0-web.ebscohost.com.oasis.unisa.ac.za/ehost/detail?vid=3&sid=ee2eb1ce-5ab2-4c19-ba67-1d8ee75abda4%40sessionmgr4003&hid=4204&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#db=ufh&AN=9605260897 [Accessed 2 December 2013].

designed with a link to a corpus and access to software for students. Kibbitzers could also be used to generate a variety of other exercises, such as cloze tests, which would be very useful for students like the participants in this study. Practically, however, this would need considerable outlay as the majority of students do not have ready access to computers or the internet.

Liu and Shaw (2001:188) add that 'vocabulary teaching would be more effective if it aimed at raising awareness of word potential so that it can be fully exploited'.This study underlines the importance of such awareness raising, revealing as it has a lack of awareness of collocational properties of simple, high-frequency verbs. The analysis of concordance lines for each of the verbs examined in this study demonstrated that students did indeed experience difficulty in producing combinations using high-frequency verbs. As Laufer and Waldman (2011:666) observe, the fact that 'use of incorrect collocations makes people sound odd but does not impair communication altogether' means that 'language accuracy' may be neglected, to the detriment of the development of 'collocational knowledge'.

This has implications for students at university; if they wish to compete in the academic milieu, students must be able to write in a way that is accurate and stylistically appropriate. As Hyland (2013:54) observes, 'English has emerged as the international language of research and scholarship'. Although these high-frequency words are often regarded as less important than academic words, and although errors in their use may not affect communication, such errors can have a cumulative effect on students' production, affecting the quality of their writing (Lee and Chen 2009:121). Nevertheless, Lee and Chen (2009) believe, like Boers et al. (2006), that simply raising learners' awareness of these collocations is not enough to help them build up a larger stock for productive use. Laufer and Waldman (2011:666), for instance, suggest that target items could be identified for learners to practise in communicative settings. Laufer and Waldman (2011) believe that because collocations such as those identified in my study are 'semantically transparent' (2011:665), they tend to be overlooked and are thus not always identified as problematic by teachers or learners. The findings of this study suggest that students would certainly benefit from such awareness raising strategies.

More than a decade after Lennon (1996) made his observations about the importance of teaching high-frequency words, Lee and Chen (2009) found in their comparison of Chinese expert and student writing an overuse of function and high-frequency words, including *make,* often when used in delexical patterns. These words were used repeatedly in their corpora in 'problematic patterns' (2009:154) such as *make an analysis/survey/communication* etc. This finding led Lee and Chen to advocate the teaching of these 'very-high-frequency common words' (2009:154) in EAP courses (and this could of course be widened to include the sort of courses focused on in the present study) rather than just academic words. Consciousness-raising activities (Lee and Chen 2009) based on the insights gained from the analysis of the corpora could be very

useful in the context described in this study. More recently, Hancioğlu et al. (2008:465) have also stressed the value of teaching words which play a scaffolding role in English. They believe that what students really need are those more general words that are used for reading and writing about academic tasks. Paquot (2010) observes too that learners need both high-frequency words and core academic vocabulary that occur across disciplines.

The findings in my study indicate less facility in the use of MWUs by writers in the Student corpora than by those in the Expert corpora, reflective perhaps of lack of depth of knowledge of high-frequency words and a limited knowledge of 'bigger', or academic words among students; this was particularly the case where the whole collocation could have been better replaced with a single-word verb and in cases where students seemed unaware of the existence of derivations. Howarth (1998b), in an explanation of the discrepancy between NS and NNS production of delexical MWUs in his findings, suggested a conscious avoidance by learners of these MWUs. I believe that in this study, overuse of *HAVE* and *TAKE* and underuse of *MAKE* in the Student Law corpus stems more from a limited knowledge of words above the 2000-word levels, an overuse of simple high-frequency verbs and only a very vague understanding that these words can be used in combination with a restricted set of nouns, and perhaps an underdeveloped understanding of collocation in general, signalling a lack of depth of knowledge of these words. Students appeared to lack what Granger (1998b) calls a sense of salience, that is, a deeper awareness of the properties of words and of which words can collocate with a particular item and which cannot.

These findings are corroborated by inadequacies in the size of students' vocabulary, indicated by generally low scores on the vocabulary test, and the fact that only some students at the 75th percentiles of these two courses had achieved mastery of levels of vocabulary above the first level of the VLT, a finding which again underscores the relationship between vocabulary knowledge and academic performance. Students recognised and could use the three verbs in question, but they were clearly not particularly adept at using them to create delexical MWUs. It seems that these students, like those in other studies, had difficulty producing collocations; for instance, Farghal and Obiedat (1995:326) found that L2 learners could not 'cope' with collocations, while Nesselhauf (2003:328) found that 'even advanced learners have considerable difficulties in the production of collocations'. This lack of understanding of collocational salience, or of restriction, is illustrated particularly in the examples of MWUs featuring the word *corruption* in the Student Law corpus: this word was frequently incorrectly collocated, most often with the verb *MAKE*.

Recommendations made by Farghal and Obiedat (1995) and Nesselhauf (2003) are particularly apposite to this study, although, like Lee and Chen (2009), Boers et al. (2006) and Laufer and Waldman (2011), their studies deal with learners of English as a foreign language. Farghal and Obiedat (1995:326) believe that 'instructors should single collocations out as the most needed and useful genre of prefabricated speech'

while Nesselhauf (2003:238) feels strongly that 'collocations do deserve a place in language teaching'. She suggests that some collocations, such as for instance those which are frequent in an academic register, should be taught explicitly. This is supported by scholars such as Martinez and Schmitt (2012) who argue for the inclusion of lists of multiword expressions such as their PHRASE List in teaching materials. This means that teachers themselves must move beyond the simple teaching of lists of words, such as word meanings, synonyms, antonyms and so on, to focus on the 'syntagmatic lexical relations' (Farghal and Obiedat 1995:326) reflected in MWUs such as those focused on in this study. Again, corpora and concordancing software provide many opportunities for gainful activities using online corpora, as suggested by Reppen (2010). These include investigating register, teaching academic language, and teaching spoken language. As far as MWUs and this study are concerned, however, word searches and the generation of KWICs can reveal vital information about the lexical relations and the way words are used in context, and in a way that is more immediate and more relevant to the learner than a vocabulary list. More specifically, when dealing with delexical MWUs students could be tasked with investigating KWICs from learner corpora where SVCs and collocational problem examples are highlighted, for instance. They could be asked to consider these and to discuss how they could be improved.

Nesselhauf (2003:238) also makes the point that it is not enough to teach merely the 'lexical elements that go together' in a collocation, but whole combinations including article, prepositions and so on should also be taught. This was particularly clear in the errors made by students in my study: in many cases other elements of the combination besides the verb and noun proved problematic. Likewise, Nesselhauf (2003:239), finding as I did that the verb caused most difficulties in verb+noun combinations, recommends that in teaching such combinations 'the focus should be on the verb'.

These are aspects of vocabulary teaching which I feel should be included in any academic support programme for students at a university such as this, where reading levels are often low and students have not had the opportunity or guidance to develop their vocabulary size or depth of word knowledge at school. Universities should be encouraged to provide academic support programmes for students who reach this level of education without the necessary vocabulary and reading skills. Programmes that extend the curriculum of three-year bachelor's degree progammes to allow for the inclusion of additional 'foundational' tuition (Boughey 2013:31) have been introduced at several universities, provided for by funding from the Department of Higher Education. In an analysis of papers submitted to a 2012 conference that focused on practitioners' understanding of the construct of academic literacy, and the approaches to literacy development emanating from such understandings, Boughey (2013:37) found that many of these papers reflected thinking that was 'far removed' from the theory in the field of New Literacy Studies characterised by the work of the likes of Street (183, 1995, cited in Boughey 2013:28), who posited in his 'ideological model' that literacy was more than simply 'the technical ability to encode and decode to and

from print'. The ideological model argues that literacy 'also involves a socially embedded disposition to interact with certain kinds of texts in certain kinds of ways' (Boughey 2013:28). Boughey (2013) advocates that if current initiatives are to make an impact on practice in South African higher education, they should be drawing on work done in this this country by academics in the field of of New Literacy Studies, and to take into account how the students they teach come from different social and cultural backgrounds. Boughey (2013:29) observes that 'home based literacies are linked to an individual's chances of accessing and succeeding in higer education regardless of the type of schooling available'. Disparities between literacies in the home environment and school and academic literacies mean that mastery of the literacies associated with education will not be supported. Gee (2008, cited in Boughey 2013:29), building on the 'ideological model', urges scholars to 'consider reading and writing as linked to values and beliefs and, importantly, to identity itself'.

In the same vein Jacobs (2013), in considering how the discipline of academic literacies has defined itself over the last two decades or so, observes a shift away from a study skills model that views literacy as a distinct cognitive skill, towards what she terms an 'academic socialisation model', which sees literacy as 'acculturating students into disciplinary discourses and focuses on disciplinary genres' (Jacobs 2013:128). She argues that the teaching of academic literacies should focus on explaining the norms and conventions of disciplines, and also allowing these to be challenged. She advocates a collaborative approach between discipline specialists and academic literacies lecturers (language specialists), which would allow the 'tacit dimension' of this knowledge of the norms and conventions to be made more explicit to students. This underscores my contention that such academic literacies courses should be an integral part of the teaching curriculum of all subjects, not just English; in the case of vocabulary in particular, all departments should take on the responsibility of helping students to develop their own vocabulary knowledge.

As mentioned above (§7.4), the fact that this study included writing from two different courses of study has pedagogical implications in the light of the above; examples from these corpora, both the Expert and the Student Literature and Law corpora, could be useful in providing models for both teachers and novice writers of how language is used in their specific field, and of the sort of errors that are typically made by learners in these particular fields. This would assist teachers in identifying expressions which might be marginally acceptable, and then to present or suggest ways in which these could be improved. Such an approach, as advocated by Hyland (2008), could thus be used to identify both combinations which occur frequently and those which might be useful for other reasons for courses such as the ones in which the students in this study were enrolled.

In addition, and reflective of the context of this study, the analysis also highlighted other areas in which students experienced difficulties, reflecting what studies on BSAE have found, for instance (Botha 2013;

Partridge 2011; Van Rooy 2006, 2013). These included aspects of verb use such as collocation, verb tense and subject-verb agreement, or concord; although language problems such as these did not form part of the focus of the analysis, writers in the student corpora showed a tendency to overuse the progressive aspect of the verb and to make errors of concord. As already observed, these are characteristics of varieties of English spoken in this country and of the English spoken by those who have a different mother-tongue; but these errors were made across the board, even by the more successful writers, leading to the impression that they may have become fossilised in the language of many writers.

These findings have pedagogical implications for a distance learning institution with hundreds of thousands of students, revealing as they do something of the nature of student writing, the demographics of the students enrolled at this university, their very diversity, their difficulties, and where teaching should focus. Without a grasp of the idiomatic and sometimes idiosyncratic way in which English words combine, a learner's writing will always sound at best foreign and at worst stilted and unnatural (Pawley and Syder 1983). One way of improving students' use of these combinations, and of English in general, is to make practical efforts to improve their reading habits through extended reading and greater exposure to texts and to make them aware of why it is so important to read extensively. Students enrolled in the two courses in this study were not required to write very much, and the quality of their writing suggested that many did not read very much either. This is not a problem unique to South Africa, but it does pose more challenges than it might in a developed country. Although ongoing efforts are being made by the DBE and bodies such as the National Library of South Africa, READ, The Molteno Project and PRAESA among others to foster reading and the development of suitable reading material for children in schools, the effects have yet to be seen among university students. As noted in Chapter 1 (§1.2), there is as yet no strongly established culture of reading in this country and in many schools very little reading occurs, and still less in homes. It is because of these factors that universities need to take on a new role, one in which it is their responsibility to place more emphasis on the value of reading and to inculcate a culture of reading among students. In the past, it was assumed that students entering university could read; today, studies have revealed that more students are arriving at universities with very low literacy levels (Meier 2011; Sondlo and Subotsky 2010; Machet and Tiemensma 2009) and that students' vocabulary, particularly their understanding of academic words, makes it difficult for them to meet the demands of reading at university (Cooper 1999, 2000). Universities need to make concerted and strenuous efforts to emphasise the high-stakes nature of reading and to use what research has shown about the importance of extended reading in order to drive the development of a reading habit. All stakeholders at universities – students, lecturers, academic support programmes and university management – should be made aware of both the intrinsic and extrinsic value of reading. All these role players need to understand that a limited vocabulary will hamper any studies students undertake, not just in English, and that reading, extensive and extended, is the best way in which to develop vocabulary. This has implications for all academic study, not just in English: the link between

reading ability and success in mathematics, for instance, is well known (Bohlmann and Pretorius 2008). Students should also be helped to appreciate the intrinsic value of reading, the simple pleasure it affords the reader.

7.6 LIMITATIONS OF THE STUDY AND SUGGESTIONS FOR FURTHER RESEARCH

There were several limitations to this study, both situational and methodological. As far as the limitations caused by the situation or research context are concerned, these were mostly beyond my control as researcher. Chief among these was the reliance on the cooperation of the students themselves in returning the questionnaire and test. In this I was bound by the rules of the university and the reality of the semester system. I was forced to communicate with students via tutorial letters and I had little control over their compliance with my request to return the material, even though I provided them with a stamped addressed envelope for this purpose. In addition, students were not in a test situation when they answered the VLT.

The original sample size was limited to those students who had returned the tests and questionnaire as requested. The return rate was 8%, which is relatively high, but as some of these students did not write the final exam, the sample was further reduced, another circumstance which was beyond my control. The sample that was collected was thus not random, determined as it was by student self-selection. This is sometimes said to skew the sample to represent more motivated students.

A further limitation was the length of texts which students wrote in the exam. Although length was stipulated, there was no guarantee that students would comply with this requirement and, in the event, Law students in particular wrote much less than I had anticipated, and they wrote only one essay where the Literature students wrote two. This was, however, mitigated to some extent by the fact that there were more Law students in the sample than Literature students. Even so, in the end the Student Law corpus was much smaller than the Student Lit corpus. These circumstances may have imposed methodological limitations on the study, by limiting the size of the specialised learner corpora in general (Student Lit corpus: 142 655 words; Student Law Corpus: 63 518 words), and the Law corpus in particular.

In the compilation of the Expert corpora I was again limited by the circumstances and could use only the texts that were available in the modules concerned, resulting in a larger Literature corpus than Law corpus (Expert Lit corpus: 144 231 words; Expert Law corpus: 47 829 words). Larger corpora containing more and longer texts might have produced richer results, and there is scope here for further research in which larger and more comprehensive corpora are collected. On the other hand, viable studies have used much smaller corpora than the ones used here, and the compilation of small, specialised corpora for research purposes is

certainly a growing trend in corpus studies (De Klerk 2006b; Flowerdew 2001; Pienaar and De Klerk 2009; Van Rooy 2006).

The size of the student corpora necessarily limited the analysis regarding the number of deviant MWUs and errors, making it difficult to obtain significant results, particularly in Phase 3 of the study, although a variety of statistical techniques were used to establish the accuracy of generalisations made about the population of the study. Larger corpora may have provided richer information on the specific nature of errors made by students and the links these have with vocabulary size. The limited size of the sample used in Phase 3 did not provide conclusive findings on the relationship between the production of MWUs and academic proficiency, although a link was suggested. There is certainly scope here for further research using larger corpora and investigating the use of other high-frequency verbs in MWUs. Such studies could also usefully compare the use of the verbs in question in their other functions; I believe it is unlikely that students would generate as many errors in their use of these verbs.

Another possible methodological limitation to the study was the omission from the questionnaire of more probing questions about language and education background, such as an indication of the actual schools students had attended. More information about the languages spoken in the educational context and in the social milieu of the individual students would also have added value to the analysis. Consequently, this provides fertile ground for further research. Future studies could look more closely at the influence of home language on the writing of students such as those in this study, particularly the influence on their use of MWUs. Also, the questionnaire should ideally in future include questions eliciting more detailed information on socioeconomic background in order to examine the effects of related factors.

Closer examination of the relationship between the use of MWUs, especially those containing high-frequency words, and reading comprehension, vocabulary levels and academic proficiency would also provide more insight into the difficulties such students have with vocabulary in general and with MWUs in particular. There is certainly scope for further research in the South African context, particularly as studies of vocabulary knowledge in African languages and the use of high-frequency words in African language corpora is scant. The perspective on depth of vocabulary knowledge in this study was a narrow one, and more work could be done in this area, for instance in the comparison of MWUs to other measures of depth, such as the Word Associates Test (WAT) used by Akbarian (2010) (see §3.5.2).

7.7 CONCLUSION

This study has taken an as yet uncommon approach to linguistic analysis by combining quantitative measures of breadth of vocabulary knowledge with a corpus approach to the measurement of a very specific aspect of depth of vocabulary knowledge. It focused on two groups of students from two courses within an English department, and found significant differences in vocabulary size between these groups of students in terms of their course of study, gender, age and language background, and also between their use of selected high-frequency verbs as compared to expert writers for the two courses.

Using examination scores as a marker of academic performance, the study has found evidence of a relationship between breadth of vocabulary knowledge and academic achievement. Although a significant relationship between vocabulary depth and academic achievement was not found when using a smaller sample of texts by those 60 students whose scores on the examination were closest to the 25th, 50th and 75th percentiles, the results did suggest a link. This is clearly an area that requires further research.

To sum up, in its attempt to throw more light on ideas about breadth and depth of vocabulary knowledge and the relationship between them, in the context of a focus on academic achievement, I believe that this study has made a meaningful contribution to vocabulary research, particularly with regard to students at tertiary level.

References

Akbarian, I. 2010. The relationship between vocabulary size and depth for ESP/EAP learners. *System*, vol. 38:391-401.

Algeo, J. 1995. Having a look at the expanded predicate. In Aarts, B. and Meyer, C.F. (Eds), *The verb in contemporary English*: *Theory and description*. Cambridge: Cambridge University Press, pp. 203–217.

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In Cowie, A.P. (Ed), Phraseology: Theory, analysis and applications. Oxford: Clarendon Press, pp. 101–122.

Altenberg, B. 2002. Causative constructions in English and Swedish. A corpus-based constructive study. In Altenberg, B. and Granger, S. (Eds). *Lexis in contrast. Corpus-based approaches*. Amsterdam/Philadelphia: John Benjamins, pp. 97–116.

Altenberg, B. and Granger, S. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics,* 22(2):173–195.

Aston, G. 1997. *Small and large corpora in language learning*. Available at: http://www.sslmit.unibo.it/~guy/wudj1.htmhttp://www.sslmit.unibo.it/~guy/wudj1.htm [Accessed 12 April 2013].

Bahns, J. 1993. Lexical collocations: a contrastive view. *ELT Journal*, January 47(1):56–63.

Bahns, J. and Eldaw, M. 1993. Should we teach EFL students collocations? *System*, 21(1):101–114.

Baker, P. 2010. Corpus methods in linguistics. In Litosseliti, L. (Ed.). 2010. *Research methods in linguistics.* London, New York: Continuum, pp. 93–113.

Bauer, L. 1993. *Manual of information to accompany the Wellington Corpus of Written New Zealand English*. Available at: http://icame.uib.no/wellman/well.htm [Accessed 26 February 2014].

Biber, D. 2006. *University language: a corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, D. 2009. A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3):275–311.

Biber, D. and Conrad, S. 2001. Quantitative corpus-based research: much more than bean counting. *TESOL Quarterly*, Summer 35(2):331–336.

Biber, D. and Finegan, E. 1991. On the exploitation of computerized corpora in variation studies. In Aijmer, K. and Altenberg, B. (Eds), *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman, pp. 204–220.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Biskup, D. 1992. L1 influence on learners renderings of English collocations: a Polish/German empirical study. In Arnaud, P.J.L. and Bejoint, H. (Eds), *Vocabulary and applied linguistics*. Basingstoke, UK: Macmillan, pp. 85–93.

Boers, F., Eykmans, J., Kappel, J., Stengers, H. and Demecheleer, M. 2006. Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Language Teaching Research*, 10(3):245–261.

Bohlmann, C. and Pretorius, E.J. 2008. Relationships between mathematics and literacy: exploring some underlying factors. *Pythagoras*, 67:42–55.

Botha, Y. 2013. Corpus evidence of anti-deletion in black South African English noun phrases. *English Today*, 29(1):16–21.

Boughey, C. 2013. What are we thinking of? A critical overview of approaches to developing academic literacy in South African higher education. *Journal for Language Teaching*, 47(2); 25−42.

British Academic Spoken English Corpus (BASE). Available at: http//www.reading.ac.uk/AcaDepts/ll/base_corpus/index.htm [Accessed 21 January 2011].

British Academic Written English (BAWE) corpus. Available at: http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/ [Accessed 21 January 2011].

CAPS. Curriculum and Assessment Policy Statement n.d. *Intermediate Phase: First Additional Language.* Department of Basic Education. Available at: http://www.education.gov.za/LinkClick.aspx?fileticket=G7iZPVv7LU8%3D&tabid=421&mid=1884 [Accessed 21 January 2014].

Coady, J., Magoto, J., Hubbard, P., Graney, J. and Mokhtari, J. 1993. High frequency vocabulary and reading proficiency in ESL readers. In Huckin, T., Haynes, M. and Coady, J. (Eds), *Second language reading and vocabulary learning*. Norwood, NJ: Ablex, pp. 217–228.

Cobb,T. *Web Vocabprofile* n.d.. Available online at http://www.lextutor.ca/vp/ [Accessed 30 April 2014], an adaptation of Heatley, A., Nation, I.S.P. and Coxhead's, A. (2002) *Range*.and *Frequency* programs. Available at http://www.victoria.ac.nz/lals/staff/paul-nation .[Accessed 3 June 2014.]

Cobb, T. 2003. Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *The Canadian Modern Language Review*, March 59(3):393–423.

Cobb, T. and Horst, M. 2001. Reading academic English: carrying learners across the lexical threshold. In Flowerdew, J. and Peacock, M. (Eds), *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press, pp. 315–329.

Coetzee, W. 2009. Language errors in the use of English by two different groups of Afrikaans first language-speakers employed by Nedbank: an analysis and possible remedy. Thesis (M Phil (General Linguistics)) University of Stellenbosch. Available at: http://hdl.handle.net/10019.1/2063 [Accesssed 5 June 2014.]

Coetzee-Van Rooy, S. 2011. Discrepancies between perceptions of English proficiency and scores on English tests: implications for teaching English in South Africa. *SAALT Journal for Language Teaching*, 45(2):151–181.

Coetzee-Van Rooy, S. 2012. Flourishing functional multilingualism: evidence from language repertoires in the Vaal Triangel region. *International Journal of the Sociology of Language*, 218:87–119.

Coetzee-Van Rooy, S. 2014. Explaining the ordinary magic of stable African multilingualism in the Vaal Triangle region in South Africa. *Journal of Mulilingual and Multicultural Development*, 35(2):121–138.

Collins, P. and Peters, P. 1988. The Australian corpus project. Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi, pp. 103–20.

Cooper, P.A. 1999. Lexis and the undergraduate: analysing needs, proficiencies and problems. Unpublished MA dissertation, University of South Africa, Pretoria.

Cooper, P.A. 2000. Academic vocabulary: putting words in academic texts in perspective. *South African Journal of Linguistics*, Supplement vol. 37:18–32.

Corpus of American Soap Operas. Brigham Young University. Available at: http://corpus.byu.edu/soap/ [Accessed 5 March 2014]

Corpus of Contemporary English (COCA). Brigham Young University. Available at: http://corpus.byu.edu/coca/ [Accessed 27 February 2014]

Corpus of Historical American English (COHA). Brigham Young University. Available at: http://byu.edu/coha [Accessed 27 February 2014]

Cowie, A.P. 1992. Multiword lexical units and communicative language teaching. In Arnaud, P.J.L. and Bejoint, H. (Eds), *Vocabulary and applied linguistics*. Basingstoke, UK: Macmillan, pp. 1–12.

Coxhead, A. 2000. A new academic wordlist. *TESOL Quarterly*, 34(2):213–238.

Cummins, J. (1999). BICS and CALP: Clarifying the distinction. Available online at: http://files.eric.ed.gov/fulltext/ED438551.pdf [Accessed 14 October 2014].

Cunningham, E.E. and Stanovich, K.E. 2001. What reading does for the mind. *Journal of Direct Instruction*, 1(2):137–149.

Cunningham, E.E. and Stanovich, K.E. 2003. Reading can make you smarter. *Principal*, November/December 83(2):34–39. Available at: http://www.naesp.org [Accessed 12 January 2014].

De Beaugrande, R. 2001. Large corpora, small corpora, and the learning of 'language'. In Ghadessy, M., Henry, A. and Roseberry, R.L. (Eds), *Small corpus studies and ELT: theory and practice*. Amsterdam: John Benjamins, pp. 3–30.

De Cock, S. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics,* 3(1):59–80.

De Cock, S. and Granger, S. 2004. High frequency words: the bête noire of lexicographers and learners alike. In Williams, G. and Vessier, S. (Eds), *Proceedings of the Eleventh Euralex International Congress*. Université de Bretagne-Sud: Lorient, pp. 233–243.

De Cock, S., Granger, S., Leech, G. and McEnery, T. 1998. An automated approach to the phrasicon of EFL learners. In Granger, S. (Ed.), *Learner English on computer*. London, New York: Longman, pp. 67–79.

De Klerk, V. 2002. Towards a corpus of black South African English. *Southern African Linguistics and Applied Language Studies*, vol. 20:25–35.

De Klerk, V. 2006a. Corpus linguistics and world Englishes: an analysis of Xhosa English. London, New York: Continuum.

De Klerk, V. 2006b. The features of 'teacher talk' in a corpus-based study of Xhosa English. *Language Matters*, 37(2):125–140.

Dörnyei, Z. 2007. *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies.* Oxford: Oxford University Press.

Erman, B. and Warren, B. 2000. The idiom principle and the open choice principle. *Text* 20(1):29–62.

Evans, S. and Green, C. 2007. Why EAP is necessary: a survey of Hong Kong tertiary students. *Journal of English for Academic Purposes,* vol. 6:3–17.

Fan, M. 2009. An exploratory study of the collocational use by ESL students: a task based approach. *System* vol. 37:110–123.

Farghal, M. and Obiedat, H. 1995. Collocations: a neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching*, vol. 33:315–331.

Fleisch, B. 2008. Primary education in crisis: why South African children underachieve in reading and mathematics. Cape Town: Juta.

Flowerdew, L. 1998. Corpus linguistic techniques applied to textlinguistics. *System*, vol. 26: 541–552.

Flowerdew, L. 2001. The exploitation of small learner corpora in EAP materials design. In Ghadessy, M., Henry, A. & Roseberry, R. (Eds), *Small corpus studies and ELT*. Amsterdam: John Benjamins, pp. 363–379.

Francis, G. 1993. A corpus-driven approach to grammar. Principles, methods and examples. In Baker, M., Francis, G. and Tognini-Bonelli, E. (Eds), *Text and technology: in honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, pp. 137–156.

Francis, W.N. and Kučera, H. 1964. Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. Providence, Rhode Island: Department of Linguistics, Brown University. Available at: http://icame.uib.no/bcm.htm [Accessed 26 February 2014].

Gilquin, G. and Gries, S.T. 2009. Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Gläser, R. 1998. The stylistic potential of phraseological units in the light of genre analysis. In Cowie, A.P. (Ed.), *Phraseology: theory, analysis and applications.* Oxford: Clarendon Press, pp. 125–143.

Granger, S. 1998a. The computer learner corpus: a versatile new source of data for SLA research. In Granger, S. (Ed.). *Learner English on computer*. London and New York: Longman, pp. 3–18.

Granger, S. 1998b. Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie, A.P. (Ed.), *Phraseology: theory, analysis and applications.* Oxford: Clarendon Press, pp. 145–160.

Granger, S. and Rayson, P. 1998. Automatic profiling of learner texts. In Granger, S. (Ed.), *Learner English on computer*. London and New York: Longman, pp. 119–131.

Granger, S. and Tyson, S. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1):17–27.

Greenbaum, L. and Mbali, C. 2002. An analysis of language problems identified in writing by low achieving first-year students, with suggestions for remediation. *South African Linguistics and Applied Language Studies,* vol. 20:233–244.

Gries, S. 2010. Corpus linguistics and theoretical linguistics. A love-hate relationship? Not necessarily … . *International Journal of Corpus Linguistics*, 15(3):327–343.

Guillot, M-N. 2005. Revisiting the methodological debate on interruptions: from measurement to classification in the annotation of data for cross-cultural research. *Pragmatics*, 15(1):25–47.

Hakuta, K. 1974. Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, 24(2):287–297.

Halliday, M.A.K. 1991. Corpus studies and probabilistic grammar. In Aijmer, K. and Altenberg, B. (Eds), *English corpus linguistics: studies in the honour of Jan Svartvik*. Longman, London, pp. 30–43.

Hancioğlu, N., Neufeld, S. and Eldridge, J. 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, vol. 27:459–479.

Harwood, N. 2002. Taking a lexical approach to teaching: principles and problems. *International Journal of Applied Linguistics*, 12(2):139–155.

Hasselgren, A. 1994. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, vol. 2:237–260.

Hazenberg, S. and Hulstijn, J.H. 1996. Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics*, 17(2):145–163.

Henning, J.G. 2006. Linking adverbials in first, second and foreign language English student writing corpora. Unpublished Master's dissertation, North-West University, Potchefstroom.

Hirsch, D. and Nation, I.S.P. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, vol. 8:689–696.

Hofland, K. and Johansson, S. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Howarth, P. 1998a. Phraseology and second language proficiency. *Applied Linguistics,* 19(1):24–44.

Howarth, P. 1998b. The phraseology of learners' academic writing. In Cowie, A.P. (Ed.), *Phraseology: theory, analysis and applications.* Oxford: Clarendon Press, pp. 161–186.

Howie, S. 2010. The relationship between early childhood backgrounds and reading achievement in low and high achieving countries in PIRLS 2006. Paper for the International Research Conference of the International Association for the Evaluation of Educational Achievement, Gothenburg, Sweden, July, 2010. Available at: http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2010/Papers/IRC2010_Howie.pdf [Accessed 8 January 2014].

Howie, S., Van Staden, S., Tshele, M., Dowse, C. and Zimmerman, L. 2012. *Progress in International Reading Literacy Study 2011 (PIRLS): South African children's reading literacy achievement.* Centre for Evaluation and Assessment, University of Pretoria. Available at: http://web.up.ac.za/sitefiles/File/publications/2013/PIRLS_2011_Report_12_Dec.PDF [Accessed 8 January 2014].

Howie, S., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit, C., Scherman, V. and Archer, E..2008. *Progress in International Reading Literacy Study 2006 (PIRLS): South African children's reading literacy achievement*. Centre for Evaluation and Assessment, University of Pretoria. Available at: http://nicspaull.files.wordpress.com/2011/04/howie-et-al-pirls-2006-sa-summary-report.pdf [Accessed 22 March 2011].

HSRC Human Sciences Reasearch Centre. 2012. Towards equity and excellence. Highlights from TIMSS 2011. The South African perspective.
Available at: http://sds.ukzn.ac.za/files/Reddy_TIMMS_seminar%20presentation.pdf [Accessed 23 July 2013].

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. 2000. Pattern grammar: a corpus-driven approach to the lexical grammar of English. Amsterdam/Philadelphia: John Benjamin.

Hyland, K. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, vol. 27:4–21.

Hyland, K. 2013. Writing in the university: education, knowledge and reputation. *Language Teaching*, 46(1): 53–70.

Hyland, K. and Tse, P. 2007. Is there an 'academic vocabulary'? *TESOL Quarterly*, June 41(2):235–253.

International Corpus of English (ICE) Available at: http://ice-corpora.net/ice/avail.htm [Accessed 18 February 2011].

International Corpus of Learner English (ICLE) Available at: http://www.uclouvain.be/en-cecl-icle.html [Accessed 5 March 2014]

Jacobs, C. 2013. Academic literacies and the question of knowledge. *Journal for Language Teaching*, 47(2): 127−140.

Jaén, M.M. 2007. A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International Journal of English Studies (IJES)*, 7(2):127–147.

Kaszubski, P. 2000. Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: a contrastive, corpus-based approach. Unpublished PhD Thesis. Poznań: Adam Mickiewicz University. Available at: http://main.amu.edu.pl/- przemka/research.html [Accessed 27 February 2013].

Kennedy, G. 1998. *An introduction to corpus linguistics*. New York: Addison Wesley Longman.

Kjellmer G. 1991. A mint of phrases. In Aijmer, K. and Altenberg, B. (Eds), *Corpus linguistics: studies in honour of Jan Svartvik.* London: Longman, pp. 111–127.

Lake, J. 2004. Using 'on the contrary': the conceptual problems for EAP students. *ELT Journal* April 58(2):137–144.

Langer, S. 2004. A formal specification of support verb constructions. In Langer, S. and Schnorbusch, D. (Eds.), *Semantik im Lexikon*. Tübingen: Narr, pp. 179-202. Available at: http://129.187.148.72/download/publikationen/05stefan_langer_dgfs.pdf [Accessed 25 May 2012].

Laufer, B. 1986. Possible changes in attitude towards vocabulary acquisition research. *International Review of Applied Linguistics,* vol. 24:69–75.

Laufer, B. 1991. The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal,* Winter, 75(4):440-448.

Laufer, B. 1992. How much lexis is necessary for reading comprehension? In Arnaud, P.J.L. and Bejoint, H. (Eds), *Vocabulary and applied linguistics*. London: Macmillan, pp. 126–132.

Laufer, B. 1994. The lexical profile of second language writing: does it change over time? *RELC Journal,* December 25(2):21–33.

Laufer, B. 1997. The lexical plight in second language reading: words you don't know, words you think you know, and words you can't guess. In Coady, J. and Huckin, T. (Eds), *Second language vocabulary acquisition: a rationale for pedagogy*. Cambridge: Cambridge University Press, pp. 20–34.

Laufer, B. 1998. The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics,* June 19(2):255–271.

Laufer, B. and Nation, I.S.P. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics,* 16(3):307–322.

Laufer, B. and Nation, I.S.P. 1999. A vocabulary-size test of controlled productive ability. *Language Testing,* January 16(1):33–51.

Laufer, B. and Paribakht, T.S. 1998. The relationship between passive and active vocabularies: effects of language learning context. *Language Learning*, September 48(4):365–391.

Laufer, B. and Waldman, T. 2011. Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning*, June 61(2):647–672.

Lee, D.Y.W. and Chen, S.X. 2009. Making a bigger deal of the smaller words: function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, vol. 18:149–165.

Lee, D. and Swales, J. 2006. A corpus-based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, vol. 25:56–75.

Leech, G. 1991. The state of the art in corpus linguistics. In Aijmer, K. and Altenberg, B. (Eds), *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman, pp. 8–29.

Lenko-Szymanska, A. 2000. How to trace the growth in learners' active vocabulary? A corpus based study. *TALC Papers*. Available at: http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm [Accessed 29 May 2012].

Lennon, P. 1996. Getting 'easy' words wrong at the advanced level. *IRAL*, 34(1):23–37.

Léon, J. 2007. Empiricism versus rationalism revisited: current corpus linguistics and Chomsky's arguments against corpus, statistics and probabilities in the 1950-1960s. In S. Matteos and P. Schmitter (Eds), *Linguistische und epistemologische Konzepte – diachron*. Munster: Nodus Publikationen, pp. 157–176. Available at: http://htl.linguist.univ-paris-diderot.fr/leon/empiricism2007.pdf [Accessed 2 February 2011].

Lewis, M.P., Simons, G.F. and Fennig, C.D. (Eds). 2013. *Ethnologue*: *Languages of the world (17th ed.):* Dallas, Texas: SIL International. Available at http: //www.ethnologue.com [Accessed 25 November 2014].

Liu, E.T.K. and Shaw, P.M. 2001. Investigating learner vocabulary: a possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *IRAL, International Review of Applied Linguistics in Language Teaching*, 39(3):171–194.

Live, A.H. 1973. The take-have phrasal in English. *Linguistics*, vol. 95:31–50.

Machet, M. and Tiemensma, L. 2009. Literacy environment in support of the development of literacy skills and voluntary reading. *Mousaion*, Special Issue 27(2):58–76.

Mahlberg, M. 2006. Lexical cohesion: corpus linguistic theory and its application in English language teaching. *International Journal of Corpus Linguistics,* 11(3):363–383.

Martinez, R. and Schmitt, N. 2012. A phrasal expressions list. *Applied Linguistics*, 33(3):299–320.

McCormick, S. 1995. *Instructing students who have literacy problems*. Englewood Cliffs, NJ: Merrill.

McEnery, T. and Gabrielatos, C. 2006. English corpus linguistics. In Aarts, B. and McMahon, A. (Eds), *The handbook of English linguistics.* Oxford: Blackwell, pp. 33–71.

McEnery, T. and Hardie, A. 2012. C*orpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Meier, C. 2011. The Foundations for Learning Campaign: helping hand or hurdle? *South African Journal of Education*, vol. 31:549–560.

Mesthrie, R. 2010. Socio-phonetics and social change: deracialisation of the GOOSE vowel in South African English. *Journal of Sociolinguistics*, 14(1):3–33.

Michigan Corpus of Academic Spoken English (MICASE),
Available online at http://quod.lib.umich.edu/m/micase/ [Accessed 7 March 2014]

Milton, J. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *EUROSLA MONOGRAPHS* SERIES 2:57–78.
Available online at: http://www.eurosla.org/monographs/EM02/Milton.pdf [Accessed 28 April 2014.]

Milton, J. and Treffers-Daller, J. 2011. Vocabulary revisited: an analysis of word knowledge of undergraduate students and it relationship with academic achievement. Paper presented at CRELLA (Centre for Research in English Language Learning and Assessment) Winter Research Seminar, 15 December 2011. Available at: http://www.beds.ac.uk/__data/assets/pdf_file/0020/191027/jenine-treffers-daller.pdf [Accessed 24 October 2013].

Milton, J. and Treffers-Daller, J. 2013. Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, 4(1):151–172.

Minow, V. 2010. Variation in the grammar of Black South African English. Frankfurt am main: Peter Lang.

Moon, R. 1992. Textual aspects of fixed expressions in learners' dictionaries. In Arnaud, P.J.L. and Bejoint, H. (Eds.), *Vocabulary and applied linguistics*. Basingstoke, UK: Macmillan, pp. 13–27.

Moon, R. 1997. Vocabulary connections: multi-word items in English. In Schmitt, N. and McCarthy, M. (Eds), *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press, pp. 40–63.

Moon, R. 1998a. *Fixed expressions and idioms in English*. Oxford: Clarendon Press.

Moon, R. 1998b. *Phrasal lexemes in English*. In Cowie, A.P. (ed.), *Phraseology: theory, analysis and applications*. Oxford: Clarendon Press, pp. 79–100.

Moon, R. 2008. Sinclair, phraseology, and lexicography. *International Journal of Lexicography,* Advance access publication 3 July, 21(3):243–254.

Morris, L. and Cobb, T. 2004. Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, vol. 32:75–87.

Mthombeni, J.S. 2010. Teacher absenteeism in schools within the Ekurhuleni South District education system. Unpublished Master's dissertation, University of Johannesburg.

Mullis, I.V.S, Kennedy, A.M., Martin, M.O. and Sainsbury, M. 2006. *PIRLS 2006 Assessment framework specifications.* Chestnut Hill: Boston College.

Murimba, S. 2005. The Southern and Eastern Africa Consortium for Monitoring Educational Quality (Sacmeq): mission, approach and projects. *Prospects*, 35(1):75–89.

Nation, I.S.P. 1983. Testing and teaching vocabulary. *Guidelines* vol. 5: 12–15.

Nation, I.S.P. 1990. *Teaching and learning vocabulary*. Massachusetts: Heinle & Heinle.

Nation, I.S.P. 1993. Vocabulary size, growth and use. In Schreuder, R. and Weltens, B. (Eds), *The bilingual lexicon*. Amsterdam/Philadelphia: John Benjamins, pp. 115–134.

Nation, I.S.P. 2001. Using small corpora to investigate learner needs: two vocabulary research tools. In Ghadessy, M., Henry, A. and Roseberry, R.L. (Eds), *Small corpus studies and ELT: theory and practice*. Amsterdam: John Benjamins, pp. 32–45.

Nation, I.S.P. 2006 How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes,* 63,1 (September/septembre): 59–82.

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

National Benchmark Tests Project (NBT) 2009. A national service to Higher Education. Available at: http://www.pmg.org.za/files/docs/090819hesa-edit.pdf http://www.nbt.ac.za [Accessed 4 June 2014.]

Nattinger, J.R. and DeCarrico, J.S. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

NEEDU (National Education Evaluation and Development Unit). 2013. NEEDU National Report 2012: The State of Literacy Teaching and Learning in the Foundation Phase (May 2013). Available online at: http://www.education.gov.za/NEEDU/tabid/860/Default.aspx [Accessed 11 April 2014]

Nel, N. and Muller, H. 2010. The impact of teachers' limited English proficiency on English second language learners in South African school. *South African Journal of Education*, 30(4):635–650.

Nel, N. and Swanepoel, E. 2010. Do the language errors of ESL teachers affect their learners? *Per Linguam*, 26(1):47–60.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2):223–242.

Nesselhauf, N. 2004. How learner corpus analysis can contribute to language teaching: a study of support verb constructions. In Aston, G., Bernardini, S. and Stewart, D. (Eds), *Corpora and language learners.* Amsterdam/Philadelphia: John Benjamins, pp. 109–124.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Paquot, M. 2010. *Academic vocabulary in learner writing: from extraction to analysis*. London: Continuum.

Partington, A. 1998. *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam/Philadelphia: John Benjamins.

Partridge, M. 2011. A comparison of lexical specificity in the communication verbs of L1 English and TE student writing. *South African Linguistics and Applied Language Studies*, 29(2):135–147.

Pawley, A. and Syder, F.H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards, J.C. and Schmidt, R.W. (Eds), *Language and communication*. London, New York: Longman, pp. 191–226.

Pennebaker, J.W. 2011. The secret life of pronouns, vol. 16:29 31 August by Magazine issue. Available at: http://www.homepage.psy.utexas.edu/HomePage/Class/Psy394V/Pennebaker/Reprints/LSA.pdf [Accessed 4 March 2013].

Penn Treebank. Available at: http://www.cis.upenn.edu/~treebank/home.html [Accessed 4 February 2014].

Pienaar, L. and De Klerk, V. 2009. Towards a corpus of South African English: corralling the sub-varieties. *Lexikos,* vol. 19:353–371.

Pravec, N. 2002. Survey of learner corpora. *ICAME Journal*, vol. 26:81–114.

Pretorius, E.J. and Mampuru, D.M. 2007. Playing football without a ball: language, reading and academic performance in a high-poverty school. *Journal of Research in Reading*, 3(1):38–58.

Pretorius, E.J. and Ribbens, R. 2005. Reading in a disadvantaged high school: issues of accomplishment, assessment and accountability. *South African Journal of Education,* 25(3):139–147.

Prinsloo, P. 2009. Discussion Document: Modelling throughput at Unisa: The key to the successful implementation of ODL. Unisa: DISA/DCLD.

Qian, D.D. 1998. Depth of vocabulary knowledge: assessing its role in adults' reading comprehension in English as a second language. PhD thesis, University of Toronto.

Qian, D. 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment performance. *Language Learning*, September 52(3):513–536.

Qing, M. 2009. *Second language vocabulary acquisition*. Bern, Switzerland: Peter Lang.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A comprehensive grammar of the English language.* London: Longman.

Rasinger, S. 2010. Quantitative methods: concepts, frameworks and issues. In Litosseliti, L. (Ed.), *Research methods in linguistics.* London, New York: Continuum, pp. 49–67.

Rayson, P. (2003). Matrix: a statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis, Lancaster University. Available at: http://eprints.lancs.ac.uk/12287/1/phd2003.pdf [Accessed 3 January 2014].

Rayson's Log-likelihood calculator. Available at: http://ucrel.lancs.ac.uk/llwizard.html. [Accessed 3 January 2014].

Read, J. 2004. Research in teaching vocabulary. *Annual Review of Applied Linguistics,* vol. 24:146–161.

Read, J. 2007. Second language vocabulary assessment: current practices and new directions. *International Journal of English Studies*, vol. 7 (2):105–125.

Reddy, V., Van der Berg, S., Janse van Rensburg, D. and Taylor, S. 2012. Educational outcomes: pathways to performance in South African high schools. *S.Afr. J. Sci.* 108(3/4), Art. #620, 8 pages. Available at: http://dx.doi.org/10.4102/sajs.v108i3/4.620 [Accessed 7 November 2013].

Renouf, A. and Sinclair, J.M. 1991. Collocational frameworks in English. In Aijmer, K. and Altenberg, B. (Eds), *English corpus linguistics: studies in the honour of Jan Svartvik*. London: Longman, pp. 128–143.

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

Reynolds, D.W. 2005. Linguistic correlates of second language literacy development: evidence from middle-grade learner essays. *Journal of Second Language Writing*, vol. 14:19–45.

Richards, J.C. and Schmidt, R. 2002. *Longman dictionary of language teaching and applied linguistics*. London: Longman.

Ringbom, H. 1998. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In Granger, S. (Ed.), *Learner English on computer*. London and New York: Longman, pp. 41–52.

Römer, U. 2005. Progressives, patterns, pedagogy: a corpus-driven approach to English progressive forms, functions, contexts and didactics. Amsterdam/Philadelphia: John Benjamins.

Rotimi, T. 2006. Hallidayan linguistics. *An Encyclopaedia of the Arts*, 4(3):157–163.

SAQMEC III The Southern and Eastern Africa Consortium for Monitoring Educational Quality. 2010–2014. Available at: http://www.sacmeq.org/ [Accessed 16 October 2014].

Scheepers, R.A. 2003. Assessing Grade 7 students' English vocabulary in different immersion contexts. Unpublished MA dissertation, University of South Africa.

Scheepers, R.A. 2006. The effects of immersion on Grade 7 learners' vocabulary size: is incidental learning of vocabulary enough? *Journal for Language Teaching*, 40(2):1–20.

Schmitt, N. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. 2008. Review article. Instructed second language vocabulary learning. *Language Teaching Research*, 12(3):329–363.

Schmitt, N. and Carter, R. 2004. Formulaic sequences in action: an introduction. In Schmitt, N. (Ed.), *Formulaic sequences*. Amsterdam/Philadelphia: John Benjamins, pp. 1–22.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing,* 18(1):55–88.

Schmitt, N. and Zimmerman, C.B. 2002. Derivative word forms: what do learners know? *TESOL Quarterly,* Summer 36(2):145–171.

Scott, M. 2001. Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In Ghadessy, M., Henry, A. and Roseberry, R.L. (Eds), *Small corpus studies and ELT: theory and practice*. Amsterdam: John Benjamins, pp. 47–67.

Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Scott, M. and Tribble, C. 2006. *Textual patterns: key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.

Shastri, S.V., Patilkulkarni, C.T. and Shastri, G.S. 1986. *Manual of information to accompany the Kolhapur Corpus of Indian English, for use with digital computers*. Available at: http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM [Accessed 26 February 2014].

Shin, D. and Nation, P. 2008. Beyond single words: the most frequent collocations in spoken English. *ELT Journal,* October 62(4):339–348.

Shirato, J. and Stapleton, P. 2007. Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, 11(4):393–412.

Simpson-Vlach, R. and Ellis, N.C. 2010. An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31(4):487–512.

Sinclair, J. McH. (Ed.) (1987): *Looking up: An Account of the COBUILD Project*. London:
Collins ELT.

Sinclair, J.M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sondlo, M. and Subotzky, G. 2010. *The challenges facing South Africa's secondary schooling system and their implications for higher education*. DISA Information and Analysis Briefing. Available at: http://www.docstoc.com/docs/132897527/Briefing-on-Secondary-education-and-its-implications-for-HE [Accessed 20 October 2013].

Spaull, N. 2012. *Poverty & Privilege: Primary School Inequality in South Africa,* Stellenbosch Economic Working Papers: 13/12, July. A working paper of the Department of Economics and the Bureau for Economic Research at the University of Stellenbosch pp 1-23.
Available at: http://www.ekon.sun.ac.za/wpapers/2012/wp132012 [Accessed 19 July 2013]

Statistics South Africa. 2011. *Census 2011 Key results,* p. 6. Available at: www.statssa.gov.za/Census2011/Products/Census_2011_Key_results.pdf [Accessed 8 January 2014].

Stein, G. 1991. The phrasal verb type 'to have a look' in modern English. *IRAL*, XXIX(1):1–29.

Stephen, D.F., Welman, J.C. and Jordaan, W.J. 2004. English language proficiency as an indicator of academic performance at a tertiary institution. *South African Journal of Human Resource Management*, 2(3):42–53.

Stubbs, M. 1986. *Educational linguistics*. Oxford New York: Basil Blackwell Ltd.

Stubbs, M. 1993. British traditions in text analysis: from Firth to Sinclair. In Baker, M., Francis, G. and Tognini-Bonelli, E. (Eds), *Text and technology: in honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, pp. 1–33.

Stubbs, M. 2001. Words and phrases: corpus studies of lexical semantics. Oxford/Massachusetts: Blackwell.

Sun, C. and Feng, G. 2009. Process approach to teaching writing applied in different teaching models. *English Language Teaching*, 2(1) March:150–155.

Taylor, C. 2008. What is corpus linguistics? What the data says. *ICAME Journal*, vol. 32:179–200.

Taylor, S., Van der Berg, S., Reddy, V. and Janse van Rensburg, D. 2011. *How well do South African schools convert Grade 8 achievement into matric outcomes?* Stellenbosch Economic Working Papers: 13/11. A working paper of the Department of Economics and the Bureau for Economic Research at the University of Stellenbosch, pp 1-47. Available at: http://www.ekon.sun.ac.za/wpapers/2011/wp132011/wp-13-2011.pdf [Accessed 5 March 2014.]

Thompson, G. and Hunston, S. 2006. System and corpus: two traditions with a common ground. In Thompson, G. and Hunston, S. (Eds), *System and corpus: exploring connections*. London: Equinox, pp. 1–14.

TIMSS. 2011. TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade. Available online at: http://timss.bc.edu/timsspirls2011/international-database.html [Accessed 16 April 2014]

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam/New York: John Benjamins.

UNISA Open Distance Learning Policy (2008) Available at: http://cm.unisa.ac.za/contents/departments/tuition_policies/docs/OpenDistanceLearning_Council3Oct08.pdf [Accessed 19 July 2013]

Van der Berg, S**.** 2008. How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation* 34(3): 145-154.
Available at: http://www.econstor.eu/bitstream/10419/32027/1/558658229.pdf [Accessed 5 March 2014]

Van Rooy, B. 2005. Expressions of modality in Black South African English. Available at: www.birmingham.ac.uk/Documents/.../vanrooyCL2005paper.doc [Accessed 11 June 2014].
Van Rooy, B. 2006. The extension of the progressive aspect in Black South African English. *World Englishes*, 25(1):37–64.

Van Rooy, B. 2011. A principled distinction between error and conventualized innovation in African Englishes. In Mukherjee, J. and Hundt, M. *Exploring second-language varieties in English and learner Englishes: bridging a paradigm gap*. Amsterdam/Philadelphia: John Benjamins, pp. 189−208.

Van Rooy, B. 2013. Corpus linguistic work on Black South African English. *English Today*, vol. 29:10–15.

Van Rooy, B. and Schäfer, L. 2003. An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. *Lancaster University Centre for Computer Corpus Research on Language Technical Papers*, vol. 16:835-844. Available at: http://www.corpus4u.org/forum/upload/forum/2005092023174960.pdf [Accessed 12 April 2013].

Van Rooy, B. and Terblanche, L. 2006. A corpus-based analysis of involved aspects of student writing. *Language Matters: Studies in the Languages of Africa*, 37(2):160–182.

Vermeer, A. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of output. *Applied Psycholinguistics,* vol. 22:217–234.

Wang, Y. and Shaw, P. 2008. Transfer and universality: collocations used in advanced Chinese and Swedish learner English. *ICAME Journal*, vol. 32:201–232.

Weinert, R. 1995. The role of formulaic language in second language acquisition: a review. *Applied Linguistics,* 16(2):180–205.

West, M.P. 1953. (1985 printing). A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology. London: Longman.

Wierzbicka, A. 1982. Why can you *have a drink* when you can't **have an eat*? *Language*, 58(4):753–799.

Wittenberg, E. and Piñango, M.M. 2011. Processing light verb constructions. *The Mental Lexicon*, 6(3):393–413.

Wray, A. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21(4):463–489.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. and Perkins, M.R. 2000. The functions of formulaic language: an integrated model. *Language and Communication*, vol. 20:1–28.

Yan, Q. 2006. A corpus-based analysis of the verb 'do' used by Chinese learners of English. *CELEA Journal*, 29(6):37–41. Available at: http://www.celea.org.cn/teic/70/70-37.pdf [Accessed 12 March 2012].

Yeh, Y. 2004.*Global perspectives on hyman language: the South African context. Programmes to increase literacy in South Africa*. Available online at http://web.stanford.edu/~jbaugh/saw/Yoo-Yoo_Literacy.html [Accessed 25 October 2014].

Xue, G and Nation, P. 1984. A university word list. *Language Learning and Communication,* 3(2):215–229.

Other websites consulted

https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-4.php
[Accessed 10 October 2014].

http://pages.uoregon.edu/stevensj/posthoc.pdf
[Accessed 10 October 2014].

http://www.beds.ac.uk/howtoapply/departments/psychology/labs/spss/Multiple_Regression
[Accessed 10 October 2014].

Rayson, P. Log Likelihood calculator. Available online at http://ucrel.lancs.ac.uk/llwizard.html [Accessed 26 October 2014].

TIMMS and PIRLS 2011. http://timss.bc.edu/timsspirls2011/international-database.html [Accessed 16 April 2014].

http://sds.ukzn.ac.za/files/Reddy_TIMMS_seminar%20presentation.pdf [Accessed 23 July 2013].

http://www.cambridge.org.br/ [Accessed 23 October 2014].

http://www.natcorp.ox.ac.uk/ [Accessed 23 October 2014].

SPSS for Psychologists. (n.d.) Palgrave. Available at: http://www.palgrave.com/pdfs/0333734718.pdf [Accessed 6 January 2014].

http://www.nlsa.ac.za/index.php/childrens-literature-rogramme (accessed 23 October 2014)

http://0-web.ebscohost.com.oasis.unisa.ac.za/ehost/detail?vid=3&sid=ee2eb1ce-5ab2-4c19-ba67-1d8ee75abda4%40sessionmgr4003&hid=4204&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#db=ufh&AN=9605260897 [Accessed 2 December 2013].

APPENDIX A

LETTER TO STUDENTS, QUESTIONNAIRE AND VOCABULARY LEVELS TEST

Study of undergraduates' vocabulary

Dear Student

***Please read the information below*.**

I am a staff member in the Department of English Studies at Unisa. At present I am engaged in doctoral research, focussing on the relationship between student vocabulary knowledge and the vocabulary of the study materials students encounter in their English studies.

To this end, I would like to find out more about students' vocabulary knowledge. This means that I am appealing to students to complete a vocabulary test. At the end of this semester, I will look at these students' examination scripts to compare their written (productive) vocabulary with the results of this test. I really hope that my findings will assist Unisa in its quest to provide open distance learning in a way which is accessible to all students.

Participation in this study is completely voluntary and you will not be disadvantaged in any way if you choose not to participate. However, I undertake to adhere strictly to the code of research ethics and to ensure your anonymity as well as the confidentiality of the results. Although the results will be used towards a Doctoral degree and also for writing papers for publication or presentations at conferences, at no time will names or identities of students be revealed. If you do decide to participate, and are interested in the scores you receive on the vocabulary test, you are very welcome to contact me and I will provide you with feedback. I can be contacted via e-mail at scheera@unisa.ac.za, or by telephone in my office on the number 012 4296914.

If you agree to participate in this study, please sign in the space provided on the following page, and then complete the vocabulary test that follows. Once you have completed it, send this tutorial letter with your signature and the completed test back to me in the envelope provided.

Thank you for helping me to make this study possible by collecting data in this way.

Yours sincerely

Ruth Scheepers

Department of English Studies

Unisa

Tel: 012 429 6914

E-mail: scheera@unisa.ac.za

Consent to participate in research study

If you would like to participate please read the following and sign where indicated.

- I understand that my name will not be used and will not appear in any report emanating from this study.

- I agree to participate in this study, but I understand that I can withdraw my agreement at any time without any obligation if I so wish.

Signature: _____ Date: _____

YYYY/MM/DD

Student number:

Now complete the test below:

Please fill in the following information before proceeding:

Student number								
Course code	**E**	**N**	**N**					

Primary language (please tick one box only)

Afrikaans		English	
IsiNdebele		IsiXhosa	
IsiZulu		Sepedi	
Sesotho		Setswana	
SiSwati		Tshivenda	
Xitsonga		Other	

Please complete the test now. It consists of **5 sections**. Please complete all sections. Please indicate the time you started the test and the time you ended in the boxes provided. Please do not use a dictionary and do not discuss you answers with anybody – I am interested in what *you* know!

I started the test at:

__ h __

VOCABULARY LEVELS TEST (Active version)

(Laufer & Nation 1995)

Complete the underlined words. The first one has been done for you.

The 2,000-word level.

Example: He was riding a *bi*cycle.

They will restore the house to its orig_____ state.

Each room has its own priv_____ bathroom.

The tot_____ number of students at the university of 12,347.

They met to ele_____ a new president.

Many companies were manufac_____ computers.

The lakes become ice-free and the snow mel_____. .

They managed to steal and hi_____ some knives.

I asked the group to inv_____ her to the party.

She shouted at him for spoi_____ the lovely evening.

You must spend less until your deb_____ are paid.

His mother looked at him with love and pri_____.

The wind roa_____ through the forest.

There was fle_____ and blood everywhere.

She earns a high sal_____ as a lawyer.

The sick child had a very high tempe_____.

The bir_____ of her first child was a difficult time.

In A.D. 636 an Arab army won a famous vic_____ over another army.

The 3,000-word level.

They need to spend less on adminis_____ and more on production.

He saw an ang_____ from Heaven.

The entire he_____ of goats was killed.

Two old men were sitting on a park ben_____ and talking.

She always showed char_____ towards those who needed help.

He has a big house in the Cape Prov_____.

Oh Harold dar_____, I am sorry. I did not mean to upset you.

Judy found herself listening to the last ec_____ of her shoes on the hard floor.

He cut three large sli_____ of bread.

He sat in the shade beneath the pa_____ trees.

He had a crazy sch_____ for perfecting the world.

They get a big thr_____ out of car-racing.

At the beginning of their journey they ecoun_____ an English couple.

Nothing illus_____ his selfishness more clearly than his behaviour to his wife.

He took the bag and tos_____ it into the bushes.

Every year she looked forward to her ann_____ holiday.

There is a defi_____ date for the wedding.

His voice was loud and sav_____, and shocked them all to silence.

The 5,000-word level

Some people find it difficult to become independent. Instead they prefer to be tied to their mother's ap_____ strings.

After finishing his degree, he entered upon a new ph_____ in his career.

The workmen cleaned up the me_____ before they left.

On Sunday, in his last se_____ in Church, the priest spoke against child abuse.

Her favourite musical instrument was a tru _____.

The building is heated by a modern heating appa _____.

He received many com_____ on his dancing skill.

People manage to buy houses by raising a mor _____ from a bank.

At the bottom of a blackboard there is a le_____ for chalk.

After falling of his bicycle, the boy was covered with bru_____ .

The child was holding a doll in her arms and hu_____ it.

We'll have to be inventive and de_____ a scheme for earning more money.

The picture looks nice; the colours bl_____ really well.

Nuts and vegetables are considered who_____ food.

The garden was full of fra_____ flowers.

Many people feel depressed and gl_____ about the future of mankind.

The University Word List level

The afflu_____ of the western world contrasts with the poverty in other parts.

The book covers a series of isolated epis_____ from history.

Farmers are introducing innova_____ that increase the productivity per worker.

They are suffering from a vitamin defic_____ .

There is a short term oscill_____ of the share index.

They had other means of acquiring wealth, pres_____ , and power.

The parts were arranged in an arrow-head configu_____ .

The learners were studying a long piece of written disco_____.

People have proposed all kinds of hypot_____ about what these things are.

The giver prefers to remain anony_____ .

The elephant is indig_____ to India.

You'll need a mini_____ deposit of R20,000.

Most towns have taken some eleme_____ civil defence precautions.

The presentation was a series of sta_____ images.

This action was necessary for the ulti_____ success of the revolution.

He had been expe_____ from school for stealing.

The lack of money depressed and frust_____ him.

The money from fruit-picking was a suppl_____ to their regular income.

The 10,000-word level

He wasn't serious about art. He just da_____ in it.

Her parents will never acq_____ to such an unsuitable marriage.

Pack the dresses so that they won't cre_____ .

Traditionally, men were expected to nu_____ women and children.

Religious people would never bl_____ against God.

The car sk_____ on the wet road.

The politician delivered an arrogant and pom_____ speech.

The Romans used to hire au_____ troops to help them in their battles.

At the funeral, the family felt depressed and mo_____ .

His pu_____ little arms and legs looked pathetic.

A vol_____ person will change moods easily.

The debate was so long and tedious that it seemed int_____.

Drink it all and leave only the dre_____.

A hungry dog will sa_____ at the smell of food.

The girl's clothes and shoes were piled up in a ju_____ on the floor.

Some monks live apart from society in total sec_____.

The enemy suffered heavy cas_____ in the battle.

When the wedding celebrations and rev_____ ended, there were plenty of drunk people everywhere.

I completed the test at:
__ h __

APPENDIX B

ETHICAL CLEARANCE

12 October 2009
College of Human Sciences

Ms Ruth Angela Scheepers
Department of English Studies, Unisa
Theo van Wijk Building Room 7-23 12 October 2009

Proposed title: THE VOCABULARY GAP: AN INVESTIGATION OF ADVANCED LEARNERS' LEXICAL DIFFICULTIES IN RELATION TO THEIR CONTENT SUBJECTS.

Principal investigator: Ms RA Scheepers

Reviewed and processed as: Class approval (see paragraph 10.7 of the UNISA Guidelines for Ethics Review)

Approval status recommended by reviewers: Approved

The Ethics Subcommittee of the College of Human Sciences has reviewed this proposal and considers the methodological, technical and ethical aspects of the proposal to be appropriate to the tasks proposed. Previously suggested additions have been affected and approval is hereby granted to the principal investigators to proceed with the study in strict accordance with the approved proposal and the ethics policy of the University of South Africa.

In addition, the principal investigator should heed the following guidelines:

- To only start this research study after obtaining informed consent
- To carry out the research according to good research practice and in an ethical manner
- To maintain the confidentiality of all data collected from or about research participants, and maintain security procedures for the protection of privacy
- To record the way in which the ethical guidelines as suggested in the proposal has been implemented in the research
- To work in close collaboration with your supervisor(s) and to notify the Subcommittee in writing immediately if any change to the study is proposed and await approval before proceeding with the proposed change
- To notify the Subcommittee in writing immediately if any adverse event occurs.

Approvals are valid for ONE academic year after which a continuation must be submitted.

Prof Kuzvinetsa Peter Dzvimbo
Deputy Executive Dean: College of Human Sciences
Tel: 012 429 4067
E-mail: dzvimkp@unisa.ac.za

UNISA university of south africa

To: Prof. RMHMoeketsi
Acting Executive Dean
College of Human Sciences
Tvw 8-08

From: Ms R Scheepers
English Studies
Tvw 6-20

Date: 12 August 2008

REQUEST FOR PERMISSION TO GATHER DATA FROM STUDENTS

Dear Prof. Moeketsi

I am registered for a doctorate in the Department of Linguistics. I am engaged in research on the relationship between the vocabulary levels of students, and the vocabulary demands of prescribed and study materials. I hope that my findings will contribute meaningfully to Unisa in terms of its open distance learning objectives.

To this end, I would like to ask students attending my department's group visits in September and October to complete a short questionnaire and a vocabulary levels test. I undertake to abide by ethical requirements by providing a letter of consent to be signed by students and by maintaining participant anonymity and confidentiality. Participation will be strictly voluntary. At the end of this year I would like to have access to these same students' examination scripts in order to analyse their productive vocabulary, using WordSmith Tools.

I hope that permission will be granted as I would like to collect this data before the end of this year.

Yours sincerely

RA Scheepers.

Ruth Scheepers
Department of English Studies

Copy: COD Prof. Z Motsa
Prof EJ Pretorius (Promoter)

Strongly supported, and looking forward to the findings of the critical investigation.

AHMoeketsi
15.08.08

RECEIVED
2008 -08- 14
OFFICE OF THE EXECUTIVE DEAN
COLLEGE OF HUMAN SCIENCES

University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

APPENDIX C

EXPERT CORPUS TEXT FILES

Files making up Expert Lit Corpus

Name	Date modified	Type	Size
101_2011_3_b	2012/02/20 10:22 ...	Text Document	33 KB
201_2011_1_b	2012/02/20 11:10 ...	Text Document	28 KB
EEDALL-3_2006_TL_301_0_Bbeginner's gu...	2012/02/20 11:41 ...	Text Document	69 KB
EEN101_02_09	2012/02/20 2:40 PM	Text Document	13 KB
ENN ALL_Group Visits CapeTown_apr10	2012/02/20 2:24 PM	Text Document	28 KB
ENN100C_2008_TL_304_B	2012/02/20 11:57 ...	Text Document	57 KB
ENN101D_202_2010_1_b	2012/02/20 12:05 ...	Text Document	25 KB
ENN101D_2010_TL_101_0_B	2012/02/20 2:52 PM	Text Document	32 KB
ENN101D_Assignment 01 Feedback_2010	2012/02/20 11:47 ...	Text Document	25 KB
ENN101D_Study Guide_2009	2012/02/24 9:23 A...	Text Document	105 KB
ENN102_103_2011_3_b	2012/02/20 11:03 ...	Text Document	42 KB
ENN102E_202_2010_1_b	2012/02/20 12:10 ...	Text Document	25 KB
ENN102E_2008_TL_103_B	2012/02/20 12:25 ...	Text Document	21 KB
ENN102E_2008_TL_104_B	2012/02/20 12:28 ...	Text Document	11 KB
ENN102E_2010_TL_101_0_B Greg	2012/02/20 2:59 PM	Text Document	37 KB
ENN102E_2011_TL_101_0_B	2012/02/20 10:59 ...	Text Document	37 KB
ENN102E_Study Guide_2006	2012/02/24 9:58 A...	Text Document	218 KB
Feedback Tutorial Letter102	2012/02/20 2:38 PM	Text Document	11 KB
Reading lecture	2012/02/20 2:22 PM	Text Document	6 KB
Self assessment_101_102_2011	2012/02/20 11:49 ...	Text Document	1 KB
TUTORIAL LETTER 10X FOR ENN102E	2012/02/20 2:18 PM	Text Document	21 KB

Files making up Expert Law Corpus

Name	Date modified	Type	Size
ENN106J_201_2009_2_b	2012/02/20 2:06 PM	Text Document	16 KB
ENN106J_201_2010_2_b	2012/02/20 2:11 PM	Text Document	15 KB
ENN106J_2011_TL_201_1_B	2012/02/20 2:14 PM	Text Document	14 KB
ENN106J_Study Guide_2008	2012/02/24 10:18 ...	Text Document	235 KB

APPENDIX D

EXAMINATION QUESTION PAPERS

The three examination question papers used in this study were scanned and have been inserted on the following pages. They are **ENN101D ENGLISH STUDIES: APPROACHING LITERATURE AND WRITING** and **ENN102E ENGLISH STUDIES: EXPLORATIONS IN READING AND MEANING**, which together made up the first-year English Literature course, and **ENN106J ENGLISH COMMUNICATION FOR LAW**, a compulsory English course for Law students. Please see pages xi to xxxiv below.

UNIVERSITY EXAMINATIONS

UNIVERSITEITSEKSAMENS

UNISA university of south africa

ENN101D

May/June 2010

ENGLISH STUDIES: APPROACHING LITERATURE AND WRITING

Duration : 2 Hours

100 Marks

EXAMINERS :

FIRST :	MR L BERMAN	DR JT PRIDMORE
	DR DW LLOYD	DR MC MARSHALL
	DR G GRAHAM-SMITH	MRS RA SCHEEPERS
	MR RN MALULEKE	
SECOND :	DR FA KALUA	PROF DC BYRNE

This paper consists of 5 pages.

Answer any TWO questions. All questions carry equal marks.

Question 1

Selves and Others

Read the following two brief extracts from *David Copperfield* by Charles Dickens and then answer the questions that follow. (The words marked by an asterisk (*) in the text are glossed at the end of the quoted passages.)

> Agnes laid aside her work, and replied, folding her hands upon one another, and looking pensively at me out of those beautiful soft eyes of hers: 'I believe he is going to enter into partnership with papa.'
>
> 'What? Uriah, That mean, fawning* fellow, worm himself into such promotion!' I cried indignantly. 'Have you made no remonstrance* about it, Agnes? Consider what a connexion it is likely to be. You must speak out. You must not allow your father to take such a mad step. You must prevent it, Agnes, while there's time.'
>
> Still looking at me, Agnes shook her head while I was speaking, with a faint smile at my warmth: and then replied:

TURN OVER

Open Rubric

'You remember our last conversation about papa? It was not long after that — not more than two or three days when he gave me the first intimation* of what I tell you. It was sad to see him struggling between his desire to represent it to me as a matter of choice on his part, and his inability to conceal that it was forced upon him. I felt very sorry.'

'Forced upon him, Agnes! Who forces it upon him?

'Uriah,' she replied, after a moment's hesitation, 'has made himself indispensable* to papa. He is subtle and watchful. He has mastered papa's weaknesses, fostered them, and taken advantage of them, unitl — to say all that I mean in a word, Trotwood, — until papa is afraid of him.'

Later in the same chapter, the narrator meets Uriah Heep. Here are his impressions:

I found Uriah Heep among the company, in a suit of black, and in deep humility. He told me, when I shook hands with him, that he was proud to be noticed by me, and that he really felt obliged* to me for my condescension*. I could have wished he had been less obliged to me, for he hovered about me in his gratitude all the rest of the evening; and whenever I said a word to Agnes, was sure, with his shadowless eyes and cadaverous* face, to be looking gauntly* down upon us from behind.

*fawning: trying to please others by paying them too much attention without being sincere

*remonstrance: protest

*intimation: sign

*indispensable: vital, very important

*obliged: in debt

*condescension: doing something that is below his social position: here, taking notice of the lowly Uriah

*cadaverous: from 'cadaver', a corpse

*gauntly: from 'gaunt', very thin; unhealthy-looking.

TURN OVER

1. From the evidence of these passages, what is the name of the narrator?
2. In one sentence, give your opinion about the character of Agnes, giving a reason for your opinion.
3. Based on the passages you have just read, what is your impression of Uriah Heep? Sum up the main features of his personality in a brief paragraph, including adjectives that describe him well.
4. Do you think Uriah Heep will turn out to be a good character, working to help others, or the opposite?
5. How did you arrive at your opinion in 4 above? Refer to and quote from the passages to substantiate your answer.

(50 marks)

Question 2

Seasons Come to Pass

Read the text of the poem below and then answer the questions that follow:

Composed upon Westminster Bridge, September 3, 1802

William Wordsworth

Earth has not anything to show more fair:
Dull would he be of soul who could pass by
A sight so touching in its majesty:
This City now doth, like a garment, wear
The beauty of the morning; silent, bare,
Ships, towers, domes, theatres, and temples lie
Open unto the fields, and to the sky;
All bright and glittering in the smokeless air.
Never did sun more beautifully steep
In his first splendour, valley, rock, or hill;
Ne'er saw I, never felt, a calm so deep!
The river glideth at his own sweet will:
Dear God! the very houses seem asleep;
And all that mighty heart is lying still!

TURN OVER

1. Briefly explain how the first three lines of the poem keep the reader in suspense as to the subject of the poem.
2. The poem is divided into an octave (the first eight lines) and a sestet (the last six lines). How does the poet's meaning follow this division?
3. Discuss the personification of London that is central to the poem, quoting from the poem to substantiate your statements.
4. Quote a line from the poem that suggests (in other words, does not say explicitly) how the city's hushed stillness when the speaker observes it contrasts with its usual noisy activity.
5. In your opinion, what is the speaker's attitude to the city at the end of the poem? Give a reason for your answer.

(50 marks)

Question 3

Heart of Darkness

The title of the novella is a metaphor for a state of mind. This is particularly so in the case of Kurtz.

Discuss this statement. Your response need not be limited to a discussion of the character of Kurtz, but may also explore other characters in the text.

(50 marks)

Question 4

The Madonna of Excelsior

Discuss the following statement:

The title, *The Madonna of Excelsior*, provides a valuable clue to Zakes Mda's

satiric critique of the injustices of the apartheid regime.

(50 marks)

Question 5

Nervous Conditions

> Stoically (Babamukuru) accepted his divinity. Filled with awe, we accepted it too. We used to marvel at how benevolent that divinity was. Babamukuru was good. We all agreed on this. More significantly, Babamukuru was right. This was why my heart swelled with gratitude as he impressed upon me the great extent of the sacrifice he had made in leaving his work to fetch me from the homestead that afternoon, impressing upon me particularly that the work that he had left to fetch me was the work that paid my school fees and bought the food that I was to eat in his house.

How are Tambu and other women dependent on Babamukuru? Using references to the novel, explain how Tambu's attitude changes towards this dependence.

(50 marks)

[TOTAL: 100 MARKS]

UNIVERSITY EXAMINATIONS

UNIVERSITEITSEKSAMENS

UNISA university of south africa

ENN102E

May/June 2010

ENGLISH STUDIES: EXPLORATIONS IN READING AND MEANING

Duration : 2 Hours

100 Marks

EXAMINERS :

FIRST :	MR L BERMAN	DR JT PRIDMORE
	DR DW LLOYD	DR MC MARSHALL
	DR G GRAHAM-SMITH	MRS RA SCHEEPERS
SECOND :	DR FA KALUA	PROF DC BYRNE

This paper consists of 4 pages.

Answer any TWO questions. All questions carry equal marks.

Question 1

Seasons Come To Pass

Read the text of the poem below, and then answer the questions that follow:

Mending Wall

Something there is that doesn't love a wall,
That sends the frozen-ground-swell under it,
And spills the upper boulders in the sun,
And makes gaps even two can pass abreast.
The work of hunters is another thing:
I have come after them and made repair
Where they have left not one stone on a stone,
But they would have the rabbit out of hiding,
To please the yelping dogs. The gaps I mean,
No one has seen them made or heard them made,
But at spring mending-time we find them there.
I let my neighbour know beyond the hill;
And on a day we meet to walk the line
And set the wall between us once again.

We keep the wall between us as we go.

TURN OVER

Open Rubric

To each the boulders that have fallen to each.
And some are loaves and some so nearly balls
We have to use a spell to make them balance:
'Stay where you are until our backs are turned!'
We wear our fingers rough with handling them.
Oh, just another kind of out-door game,
One on a side. It comes to little more:
There where it is we do not need the wall:
He is all pine and I am apple orchard.
My apple trees will never get across
And eat the cones under his pines, I tell him.
He only says, 'Good fences make good neighbours'.
Spring is the mischief in me, and I wonder
If I could put a notion in his head:
'Why do they make good neighbours? Isn't it
Where there are cows?
But here there are no cows.
Before I built a wall I'd ask to know
What I was walling in or walling out,
And to whom I was like to give offence.
Something there is that doesn't love a wall,
That wants it down.' I could say 'Elves' to him,
But it's not elves exactly, and I'd rather
He said it for himself. I see him there
Bringing a stone grasped firmly by the top
In each hand, like an old-stone savage armed.
He moves in darkness as it seems to me~
Not of woods only and the shade of trees.
He will not go behind his father's saying,
And he likes having thought of it so well
He says again, "Good fences make good neighbours."

Robert Frost

1. How would you describe the speaker and his tone in the poem? How would you describe the neighbour?
2. What do 'walls' come to represent in the poem?
3. The word 'mending' appears in the title of the poem, as well as in the action that takes place throughout the poem. What is the meaning and significance of this word in the poem?
4. The gaps in the wall serve as a metaphor for the relationship between the

5. neighbours. Briefly discuss this statement, quoting from the poem to substantiate your statements.

6. Throughout the poem, Frost plays with form to convey meaning. Give two examples from the poem where Frost uses changes in rhythm, visual gaps or syntax to convey the 'gaps' referred to in the previous question.

7. In your opinion, has the speaker closed the 'gaps' in the relationship between him and his neighbour by the end of the poem? Give a reason for your answer.

(50 marks)

Question 2

The Great Gatsby

Jay Gatsby dies at the end of the novel. Write an essay in which you discuss whom you consider to be most responsible for Gatsby's death. Is it Tom? Daisy? Myrtle? Gatsby himself? In your essay you should give reasons why or why not each character is implicated in Gatsby's death.

(50 marks)

Question 3

The Merchant of Venice

It has been said that Shylock deserves his punishment because he was motivated by the desire for revenge.

Read the following extract, and then debate the validity of this claim:

Salarino: Why, I am sure, if he forfeit, thou wilt not take his flesh. What's that good for?

Shylock: To bait fish withal. If it will feed nothing else, it will feed my revenge. He hath disgraced me...scorned my nation... . And what's his reason? I am a Jew.

TURN OVER

Hath not a Jew eyes? ...feed with the same food, hurt with the same weapons, subject to the same diseases...as a Christian is? If you prick us do we not bleed?...And if you wrong us, shall we not revenge? If a Jew wrong a Christian, what is his humility? Revenge. If a Christian wrong a Jew, what should his sufferance be by Christian example? Why, revenge.

[Act 3. Scene 1. ll 48-67]

(50 marks)

Question 4

Disgrace

From the beginning to the end of the novel, David Lurie suffers one devastating humiliation after another. He loses his job and his reputation. He is forced to flee Cape Town to live with his daughter on her smallholding in the country. There he is beaten and burned and trapped helplessly in the bathroom while his daughter is raped. Finally, he ends up ferrying dead dogs to the incinerator.

Write an essay in which you discuss whether there is a meaning or purpose in his suffering; whether he is, in some way better off at the end of the novel than he was at the beginning, and how he has changed. You must refer to the novel in detail to substantiate what you say.

(50 marks)

[TOTAL: 100 MARKS]

UNIVERSITY EXAMINATIONS

UNIVERSITEITSEKSAMENS

UNISA university of south africa

ENN106J (472184) May/June 2010

ENGLISH COMMUNICATION FOR LAW (ENGLISH 106)

Duration : 2 Hours 100 Marks

EXAMINERS :
FIRST : MR RA MUSVOTO MR BM NCHINDILA
SECOND : MS LM MASEHELA PROF B SPENCER

This paper consists of 14 pages plus instructions for completion of a mark reading sheet.

INSTRUCTIONS

This examination paper consists of TWO sections. Make sure that you answer them both. Keep track of your time and spend an hour on each section. Record the answers to section A on the mark-reading sheet. Write your answer to Section B in the examination booklet provided.

SECTION A: Reading Comprehension

Read the case Minister of Law and Order v Milne on pages 8-14 and then answer questions 1 to 25 below.

Use the following statements to answer questions 1-3

a. It is in italics.
b. It ends with the word *postea* and a date.
c. It is in full sentences.
d. It is made up of cryptic notes.
e. It contains a detailed narrative of the events.
f. It concludes with the judge's decision.
g. It includes a full discussion of the principles pertaining to the case.
h. It lists the reported cases.
i. It provides details of the main players in the current case.
j. It provides a brief summary of the case.

[TURN OVER]

Open Rubric

Question 1
Which of the above statements apply to the headnotes section of the case?
1. a d f j
2. a d f g
3. d j
4. a d j h

Question 2
Which of the above statements apply to the Annotations section of the case?
1. b c d i
2. a b h i
3. b h j
4. b h i

Question 3
Which of the above statements apply to the section of the case containing the Judge's words?
1. c e g f
2. c e f g
3. c f i j
4. a b c

Question 4
S 49 of the Criminal procedure Act 51 of 1997 (see F p. 292) is not relevant here because
1. this section authorises the use of force to prevent someone from resisting arrest.
2. this section does not authorise the use of force to prevent someone from resisting arrest.
3. the driver was drunk but not resisting arrest.
4. Sergeant Assor shot the driver because he was resisting arrest.

Question 5
The headnotes include the phrase *'Test an objective one'*. Choose the correct meaning for *objective* in this context.
1. What a reasonable man would have done in the circumstances.
2. The risk of death or injury must be imminent.
3. In keeping with contemporary notions of the value of life.
4. The risk of death or injury must be real.

Question 6
'Harm caused by negligence also unlawful and may be defended against' (Headnotes at C – D).
Choose the best interpretation of this sentence in the context of this case.
1. The driver's negligence in failing to realise that he was on the wrong side of the road could have caused harm.
2. The policeman used excessive force to stop the driver.
3. The policeman had a duty to prevent the driver's negligence from causing more harm.
4. The policeman caused harm by neglecting to ascertain that his actions were excessive in the circumstances.

[TURN OVER]

Question 7
Who is being held legally responsible for the death of the driver?
1. Constable Stander
2. Sergeant Assor
3. The Minister of Law and Order
4. All three of the above

Question 8
The first thing the two policemen saw on the night was a
1. yellow Volkswagen Golf.
2. Ford Laser.
3. vehicle marked with the insignia of the flying squad.
4. flashing blue light.

Question 9
The following events led up to the shooting on the night of the incident. Arrange them in the correct sequence:
a. the speeding driver went through two intersections without stopping
b. the policemen turned on their siren and gave chase
c. the policemen radioed for the registration of the car
d. the policemen used the loudhailer to shout at the driver
e. the policemen saw a car travelling against the flow of traffic

1. e d b c a
2. e b d a c
3. e b d c a
4. a e c b d

Question 10
The following events led to the death of Mr Milne. Arrange them in the correct sequence:
a. Sergeant Assor fired shots at the car, resting his gun on the wing mirror
b. the speeding driver took evasive action
c. the police car spun out of control on the side of the road
d. shots shattered the rear window of the speeding car and it came to a stop
e. Constable Stander fired a shot into the ground

1. e b a d c
2. b e a d c
3. b c e a d
4. e a b d c

Question 11
Sergeant Stander fired at the ground in order to
1. check whether the car had been stolen.
2. get the car off the road.
3. get the driver's attention.
4. puncture a tyre.

[TURN OVER]

Question 12
Sergeant Assor fired at the car in order to
1. check whether the car had been stolen.
2. get the car off the road.
3. get the driver's attention.
4. kill the driver.

Question 13
On the night of the incident the deceased had been
1. driving in a stolen car on the wrong side of the dual carriageway.
2. probably too drunk to notice the police car with its blue light coming towards him.
3. returning home in a car he had reported stolen.
4. driving in a stolen car under the influence of alcohol on the wrong side of the dual carriageway.

Question 14
The deceased was guilty of three of the following offences. Which is the exception?
1. driving under the influence of alcohol
2. driving a stolen vehicle
3. driving on the wrong side of a dual carriageway
4. failing to notify police that his vehicle had been recovered

Question 15
Sergeant Assor believed it was his duty to shoot the driver in order to prevent a possible fatal collision. Judge Nugent believed three of the following. Which option did the judge NOT believe?
1. Killing the driver was not an excessive response to the risk.
2. A collision was not certain and therefore the killing was not justified.
3. There was a reasonable prospect that a collision would have been avoided.
4. A collision might not have resulted in serious harm to life and limb.

Question 16
The best word to replace *quantum* (at D – E on p. 292) in this context is
1. quantity.
2. quotient.
3. amount.
4. portion.

Question 17
Match the legal principles in *a* to *d* with the names of the cases below them:
a. Killing in self-defence is justified if the person concerned had been unlawfully attacked and thought he was in danger of death or injury.
b. It is lawful for any person to use a reasonable degree of force for his or any other person's protection against any unlawful use of force. Force is not reasonable if it is either unnecessary or disproportionate to the evil to be prevented.
c. Once there is some risk of death or injury, resort may necessarily be had to lethal force because that is the only means available to repel the risk.
d. The law set bounds to the right to defence by the rule that the force used must not be out of proportion to the apparent urgency of the situation.

[TURN OVER]

A. *Pollock on Torts* 15th ed
B. *R v Attwood* 1946
C. *Nrsomi v Minister of Law and Order* 1990
D. *Salmond and Heuston* on Torts 19th ed.

1. aB bD cC dA
2. aD bC cB dA
3. aB bA cD dB
4. aB bD cA dC

Question 18
The phrase *'as readily'* (C - D in headnotes and E on p. 294) implies there is a comparison between a case of negligence and a case in which
1. there is only a possibility of harm being caused.
2. the risk of death or injury is `real and imminent'.
3. harm to life and limb is deliberately threatened.
4. there is no possibility of harm being caused.

Question 19
What specific issue does the current court deal with?
1. action for damages
2. question of liability
3. risk of death or serious injury
4. self defence

Question 20
Who is the appellant in this case?
1. Sergeant Assor
2. the Minister of Law and Order
3. Mrs Milne
4. Constable Stander

Question 21
Who represented the respondent in the first case?
1. GEAN Barlow
2. B Roux
3. Jose da Silva
4. the State Attorney

Question 22

Which of the following formed part of the judge's conclusions?

a. There was a reasonable prospect that a collision would occur.
b. Even though another vehicle may have been approaching, death or serious injury would not necessarily have occurred.
c. A real risk of death or serious injury was imminent.
d. Given the danger that the speeding car presented, Sergeant Assor was justified in killing the driver.
e. Given the danger that the speeding car presented, Sergeant Assor was not justified in killing the driver.
f. Even though another vehicle may not have been approaching, death or serious injury would necessarily have occurred.

1. a b e
2. a b d
3. b c d f
4. a c d f

Question 23

The court *a quo*

1. dealt with the question of both liability and damages.
2. found that the policeman was liable to the respondent.
3. found that the Minister of Law and Order was liable to the respondent.
4. All of the above.

Question 24

Sergeant Assor claimed that he would be liable for the consequences of an accident if he had not shot Mr Milne. The judge disputed this

1. by referring to *Minister van Polisie v Ewels* 1975.
2. by saying that SA law rarely recognises liability for something one has omitted to do.
3. by saying that Sergeant Assor had not himself created the danger.
4. All of the above.

Question 25

This case was heard in the

1. Appeal Court by Judge Goldblatt.
2. Witwatersrand Local Division by Judge Nugent.
3. Witwatersrand Local Division by Judges Nugent, Marais and Hussain.
4. Appeal Court by Judges Nugent, Marais and Hussain.

TOTAL SECTION A: 25 x 2 = 50 marks

[TURN OVER]

SECTION B

Essay Question

Choose only one of the following questions.

Write an argumentative essay of about one and a half to two pages (450 words) on ONE of the following topics.

1. *Immigration has contributed to a rise in international crime.* Discuss this statement by giving examples and show whether or not you agree with it.

2. *In some countries, the laws are punitive and not corrective.* Discuss this statement by giving examples and show whether or not you agree with it.

3. *Corruption is mainly practised by the rich.* Discuss this statement by giving examples and show whether or not you agree with it.

TOTAL SECTION B: 50 marks

TOTAL FOR PAPER: 100 marks

[TURN OVER]

MINISTER OF LAW AND ORDER v MILNE

WITWATERSRAND LOCAL DIVISION

MARAIS J, NUGENT and HUSSAIN AJ

1997 March 20, 27 Case No A5025/96

Delict – Liability for – Wrongful and negligent causing of death – Defences – Killing in defence of another – Force used must not only be necessary, but also not excessive-Where there is some risk of death or injury, resort not necessarily to be had to lethal force because such only means available to repel risk – Risk of death or serious injury required to be real and imminent – Test an objective one – Harm caused by negligence also unlawful and may be defended against – But nature of conduct in cases of negligence such that it ought not to be as readily assumed that apparent threat will manifest itself in harm – Motorist driving in wrong direction on carriageway of dual carriageway – Motorist driving at high speed – Policeman following motorist and attempting unsuccessfully to shoot tyres of motorist's vehicle – Thereafter policeman shooting in direction of motorist and killing them – Main concern of policemen was to get motorist off the road because of danger on oncoming traffic – Real risk that death or serious injury imminent not established – Killing of deceased not justified – Minister liable.

In an action for damages for the wrongful and negligent causing death, for the defence of killing in defence of another to succeed the force which was used must not only be necessary, but must also not be excessive. These are separate and distinct requirements. It ought not to be thought that, once there is some risk of death or injury, resort may necessarily be had to lethal force merely because that is the only means available to repel the risk. (At 293C.)

The risk of death or serious injury must be real and imminent in order to justify homicide and this requirement is in keeping with contemporary notions of the value to be attached to human life. The test is, of course, an objective one. What must be asked is whether a reasonable man in the position of the actor would have considered that there was a real risk that death or serious injury was imminent. (At 294B/C-D.)

While harm caused by negligence is also unlawful and may be defended against, the very nature of the conduct which is in issue in cases of negligence is such that it ought not to be as readily assumed that the apparent threat will in due course manifest itself in harm. (At 294D/E-E/F.)

In the present case, an action for damages arising out of death of the respondent's husband who had been shot and killed by a policeman in the employ of the appellant, it appeared that the deceased had been driving at high speed in the wrong direction on a carriageway of a dual carriageway. Two policemen set off in a police car in pursuit of him and, when they caught up with him, they tried, using a loudhailer, to call upon him to stop. There was no reaction from the deceased. When it became clear that the deceased did not intend to stop, one of the policemen started to shoot at the wheels of the deceased's vehicle in order to try and get the deceased's vehicle off the road because of the danger which it posed to oncoming traffic. This had no apparent effect, however, and the policeman then

[TURN OVER]

started firing in the direction of the driver. The policeman at first directed his shots low down but he gradually aimed higher as he continued firing. The deceased's vehicle then came to a stop at the side of the road. The driver of the vehicle was found dead, slumped in the seat of the car. An action for damages was instituted in a Local Division. On trial only the question of liability was dealt with. It was held that the killing of the deceased was not justified and that the appellant (defendant) was accordingly liable. In an appeal to a Full Bench,

Held, applying the principles set out above, that the trial Court had correctly concluded that a real risk of death or serious injury resulting from the deceased's conduct had not been presented itself and that the killing of the deceased was accordingly not justified. (At 295A/B.) Appeal dismissed.

Annotations:

Reported cases

Minister van Polisie v Ewels 1975 (3) SA 590 (A): distinguished
Ntsomi v Minister of Law and Order 1990 (1) SA 152 (C): dictum at 530D not approved
R v Atwood 1946 AD 331: applied
R v Molife 1940 AD 202: applied
R v Patel 1959 (3) SA 121 (A): applied.

Appeal to a Full Court from a decision of a single Judge (Goldblatt J) in the Witwatersrand Local Division. The facts appear from the judgement of Nugent J.

B Roux for the applicant.
G E A N Barlow for the respondent.
Cu adv vult.

Postea (March 27).

Nugent J: This appeal concerns the circumstances in which homicide may be justified in self-defence or in defence of another.

On the night of Friday, 19 July 1991, two members of the South African Police were in duty in a yellow Volkswagen Golf motor vehicle which was marked with the insignia of the flying squad. The vehicle was fitted with a radio, a siren and a loudhailer. Mounted on the inner side of the rear window was a blue light which would flash when illuminated. The driver of the vehicle was Sergeant Assor. Constable Stander was in the passenger seat. Their vehicle was stationary alongside the duel carriageway between Modderfontein and Johannesburg, where they were observing the passing traffic.

The two carriageways are separated by a strip of land of considerable width, which in parts is open grassland and in parts is planted with groves of trees. Each carriageway has two traffic lanes. At intersections the carriageway widens to provide a third lane for turning traffic. The road is unlit.

At about 10:00 pm the two policemen were astonished to see a Ford Laser motor vehicle pass them by at considerable speed, travelling in the direction of Johannesburg. What caused their astonishment was that the vehicle was travelling on the wrong carriageway, against the flow of traffic.

[TURN OVER]

They immediately switched on the siren and the blue light in their vehicle and set off in pursuit. They quickly caught up behind the Ford, and Constable Stander used the loudhailer to call upon the driver to stop. The sound of the siren was interrupted only while the loudhailer was being used. By this time the policemen were travelling at about 160kph immediately behind the Ford, on the inner traffic lane of the incorrect carriageway. There appeared to be no reaction from the driver. Constable Stander made radio contact with the control station of the flying squad, and conveyed the registration number of the Ford. The two policemen were informed that the vehicle was listed on the police computer as having been stolen. Training and experience had taught Sergeant Assor that he should not draw alongside or ahead of a vehicle in a situation like this, to avoid the danger of being fired upon or forced off the road.

The vehicle passed through two intersections controlled by traffic lights without stopping. The police vehicle was still following behind the Ford, with the siren sounding, but still there was no indication that the driver intended stopping. At first the road was relatively straight. Sergeant Assor said that from time to time they were passed by other vehicles approaching from the opposite direction. The traffic could not have been too heavy though, because they travelled a considerable distance without mishap.

In due course they approached a point at which the road curved to the right. Shortly before they reached it, two vehicles emerged from the curve, one travelling in each traffic lane. The driver of the Ford took evasive action, as did Sergeant Assor, who moved off the road to his right, and the police car spun out of control on the strip of land separating the two carriageways.

Sergeant Assor regained control of the police vehicle and resumed the chase. When he again caught up with the Ford, it was still travelling in the same traffic lane, but at a much reduced speed. Sergeant Assor estimated that it was travelling at about 80 kph. The police vehicle was travelling at the same speed about three to four time metres behind the Ford. Two witnesses who were travelling in a vehicle which passed in the opposite direction, probably at about this time, estimated that their car speed was about 50 to 60 kph.

At this stage Constable Stander took out his pistol and fired a shot into the ground. Sergeant Assor said their main concern was to get the Ford off the road because of the danger which it posed to oncoming traffic, and he told Constable Stander to fire at the wheels of the Ford.

The vehicles were by now approaching another blind curve in the road. Constable Stander commenced firing at the wheels of the Ford. Sergeant Stander also drew his pistol. After about six shots had been fired by Constable Stander, without apparent effect, Sergeant Assor rested his pistol on the wing mirror of the vehicle and started in the direction of the driver of the Ford. He said that at first his shots were directed low down, but gradually moved higher as he continued firing. The last few shots shattered the rear window of the Ford, and it came to a stop on the side of the road. By then he had fired fifteen shots, ceasing only when the magazine was empty.

[TURN OVER]

A The driver of the Ford was found dead, slumped in the seat of the car. One bullet had penetrated his pelvis, and another had penetrated the back of his neck and entered the brain. He was in fact the owner of the vehicle which he was driving. Some six months earlier he had reported that he had
B subsequently recovered it. Clearly he was drunk at the time the incident occurred. The concentration of alcohol in his blood was 0,29 mg per 100 ml. According to the evidence he had commenced drinking with friends at work early that afternoon, and he had then repaired to a bowling club where he continued drinking. The incident occurred shortly after he left the bowling club.

C The trial Court found that in his intoxicated state the deceased probably drove onto the carriageway on the wrong side without realising that it was a dual carriageway. Whether his intoxication also rendered him unaware that he was being pursued, or whether he was deliberately attempting to avoid the police, is not really material.

D The deceased was the husband of the respondent and the father of their two minor children. The respondent sued the appellant for the loss suffered by her and the two minor children as a result of her husband's death. The *quantum* of her claim was held over for later determination, and the Court *a quo* dealt only with the question of liability. It held that the killing of the
E deceased was not justified and that the appellant was accordingly liable to the respondent. The appellant now appeals against that decision.

Both in the Court below and in argument before us it was accepted, correctly in my view, that s 49 of the Criminal Procedure Act 51 of 1977 has no application in the present case. That section authorises the use of
F force which is reasonably necessary to prevent a person from avoiding arrest. Although the police had not been able to stop the deceased, there was no immediate danger that he would evade them entirely. Nor indeed was that what prompted Sergeant Assor to shoot the deceased. He was shot rather because of the danger which his conduct created for other users of the road.

G I think it is clear that Sergeant Assor fired at the deceased with full knowledge that he was likely to be killed, if not with that consciously in mind. He said that his purpose in doing so was to prevent the vehicle from entering the blind curve which it was approaching. He believed that if one or more vehicles were to be approaching in the opposite direction when the
H Ford entered the curve, there was likely to be a collision with fatal consequences. The question then is whether the presence of that risk justified killing the deceased.

It is well established in our law that there are circumstances in which it will be permissible to kill a person in self-defence or in defence of another. Although the problem has generally arisen in criminal cases, the principle
I is equally applicable in defence of a civil claim, save that the incidence and nature of the *onus* differs.

The approach which is taken by our laws was set out in *R v Molife* 1940 AD 202 at 204 and *R v Attwood* 1946 AD 331 at 340 (see too *R v Patel* 1959 (3) SA 121 (A) at 123A). In *Attwood's* case Watermeyer CJ said that
J homicide in self-defence is justified if the person concerned

[TURN OVER]

'had been unlawfully attacked and had reasonable grounds for thinking that he was in danger of death or serious injury, that the means he used were not excessive in relation to the danger, and that the means he used were the only or least dangerous means whereby he could have avoided the danger'.

In *Salmon and Heuston on Torts* 19th ed at 142 the following is said:

'It is lawful for any person to use a reasonable degree of force for the protection of himself or any other person against unlawful use of force... Force is not reasonable if it is either (i) unnecessary – ie greater than is requisite for the purpose – or (ii) disproportionate to the evil to be prevented.'

For the defence to succeed then, the force which was used must not only be necessary, but must also be not excessive. These are separate and distinct requirements. It ought not to be thought that, once there is some risk of death or injury, resort may necessarily be had to lethal force merely because that is the only means available to repel the risk. To the extent that this may be what is suggested by the *dictum in Ntsomi v Minister of Law and Order 1990* (1) SA 512 (C) at 530D, in my view, it does not accord with the decisions to which I have referred.

The authors of *Salmon and Heuston (supra)* add the following qualification to the passage which I have cited:

'In order that it may be deemed reasonable within the meaning of this rule, it is not enough that the force was not more than was necessary for the purpose in hand. For even though not more than necessary it may be unreasonably disproportionate to the nature of the evil sought to be avoided.'

So, too, in *Pollock on Torts* 15th ed at 123 the following is said, which seems to me to accord with our law:

'The law set bounds to (the right of defence) by the rule that the force employed must not be out of proportion to the apparent urgency of the situation...'

The distinction between the two requirements which I have referred to is particularly relevant in the present case. Certainly there was some risk that another vehicle would be approaching when the Ford entered the curve, and that a fatal collision could occur. I have accepted too that the only means by which the vehicle could have been prevented from entering the curve was by killing the driver. The question which still remains is whether this was an excessive response to the risk.

In support of his argument that Sergeant Assor's response was not excessive, the appellant's counsel went so far as to submit that he was not only justified in shooting but that he had a duty to do so, and that he would have exposed himself to liability for the consequences of a collision if he had not done so. I doubt that this is so. He had done nothing to create danger, and our law seldom recognises liability mere omissions. Although in *Minister van Polisie v Ewels* 1975 (3) SA 590 (A) at 597 Rumpff CJ said that there are circumstances in which a mere omission will be actionable, in my view, the circumstances in which that will apply do not necessarily coincide with the circumstances which are required to exist in order for an otherwise unlawful act to be justified. However, even if I am wrong in that respect, to ask whether Sergeant Assor had a duty to act as he did would beg the question which we are called upon to decide.

[TURN OVER]

The law does not provide a ready answer to the question whether the risk of death or serious injury is sufficient in any particular case to justify killing in order to avoid it. In *R v Patel (supra)* Holmes JA quoted with approval a passage from Gardiner and Lansdown in which it was said that the danger should be 'great'. *Pollock (supra)* says that an 'honest and reasonable belief of immediate danger' is required. The *American Restatement*, dealing with a danger arising from negligence (vol 1 para 66), requires a reasonable belief that the offensive conduct 'will cause' death or serious bodily harm.

In the Court *a quo* Goldblatt J was of the view that the risk of death or serious injury must be 'real and imminent' in order to justify homicide, which, in my view, aptly summarises the effect of the authorities and is in keeping with contemporary notions of the value to be attached to human life. The test is, of course, an objective one. What must be asked is whether a reasonable man in the position of the actor would have considered that there was a real risk that death or serious injury was imminent.

The problem most often arises where harm to life or limb is deliberately threatened, and in such cases it might readily be assumed that the threat will be carried to its conclusion if it is not interrupted. While harm caused by negligence is also unlawful and may be defended against (*Restatement* vol 1 para 66; *Prosser on Torts* 4th ed para 19; *Fleming on Torts* 8th ed at 84 fn 67; see, too, Neethling *et al Law of Delict* 2nd ed at 71), the very nature of the conduct which is in issue in case of negligence is such that it ought not to be as readily assumed that the apparent threat will in due course manifest itself in harm.

In the present case the Court *a quo* found that there was only a possibility of harm being caused by the deceased and that the action of Sergeant Assor was excessive. The argument on behalf of the appellant centred on this finding, and more particularly on the Court's finding that there was no more than a possibility that there may have been oncoming traffic in the same lane as the deceased. It was submitted that in view of the evidence of the traffic conditions which then prevailed the learned Judge ought to have found that it was probable that there would have been oncoming traffic in the same lane as the deceased as he entered the curve.

When Courts deal with matters of probability they are dealing with matters which might be expected to occur in the normal course of human affairs, and not with the statistical probability that an event will occur. In my view, the evidence is not sufficient to determine what chance there was that another vehicle was in that traffic lane at the material moment. In any event, the question which we were required to determine depends not on what the actual risk was, but rather on how it would have been assessed by a reasonable man in the position of Sergeant Assor (*Arwood's case supra*).

But even if it is accepted that the risk of another vehicle being present at that time was a real one, and I do not think that on the evidence it can be placed higher than that, death or serious injury would not necessarily have followed as a matter of course. As pointed out by the learned Judge in the Court *a quo*. there was a reasonable prospect that a collision would

[TURN OVER]

in any event have been avoided, bearing in mind that a collision had indeed been avoided earlier at a higher speed: and if there was indeed a collision it might not have resulted in serious injury to life and limb. A

With all these imponderables in mind, the learned Judge concluded that a real risk that death or serious injury was imminent had not presented itself, and that the killing of the deceased was accordingly not justified. I agree with that conclusion, and I would accordingly dismiss the appeal with costs. B

Marais J and Hussain AJ concurred.

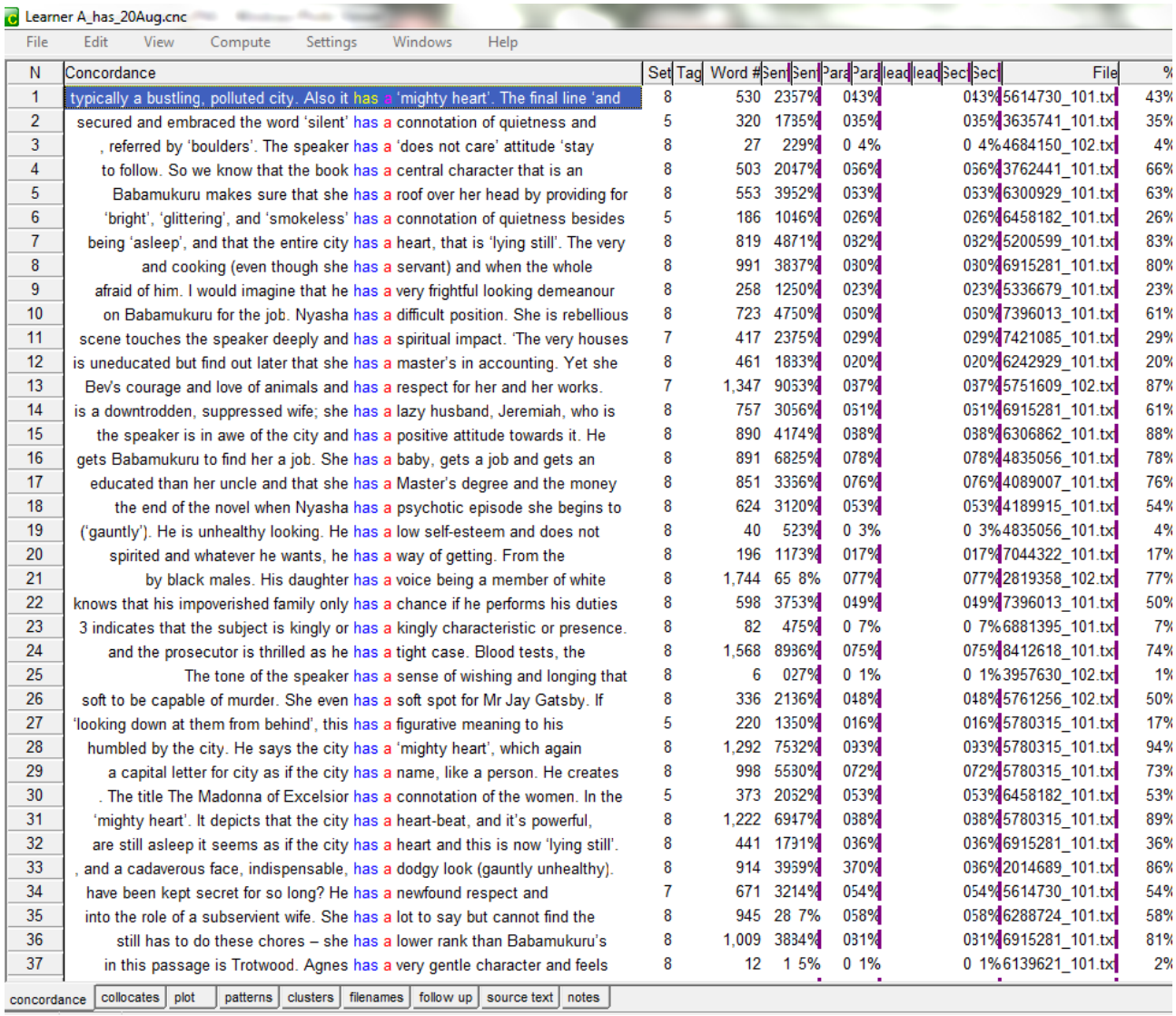
Appellant's Attorney: *State Attorney*. Respondent's Attorney: *Jose da Silva*. C

APPENDIX E

SAMPLE CONCORDANCE PAGES

Below are screenshots of concordance lines generated by WST, one page for each verb:

Student Lit *has*



Learner A_has_20Aug.cnc

File Edit View Compute Settings Windows Help

N	Concordance	Set	Tag	Word #	Sen	Sen	Para	Para	lead	lead	Sect	Sect	File	%
1	typically a bustling, polluted city. Also it has a 'mighty heart'. The final line 'and	8		530	2357%		043%				043%		5614730_101.txt	43%
2	secured and embraced the word 'silent' has a connotation of quietness and	5		320	1735%		035%				035%		3635741_101.txt	35%
3	, referred by 'boulders'. The speaker has a 'does not care' attitude 'stay	8		27	229%		0 4%				0 4%		4684150_102.txt	4%
4	to follow. So we know that the book has a central character that is an	8		503	2047%		036%				036%		3762441_101.txt	66%
5	Babamukuru makes sure that she has a roof over her head by providing for	8		553	3952%		053%				053%		6300929_101.txt	63%
6	'bright', 'glittering', and 'smokeless' has a connotation of quietness besides	5		186	1046%		026%				026%		6458182_101.txt	26%
7	being 'asleep', and that the entire city has a heart, that is 'lying still'. The very	8		819	4871%		032%				032%		5200599_101.txt	83%
8	and cooking (even though she has a servant) and when the whole	8		991	3837%		030%				030%		6915281_101.txt	80%
9	afraid of him. I would imagine that he has a very frightful looking demeanour	8		258	1250%		023%				023%		5336679_101.txt	23%
10	on Babamukuru for the job. Nyasha has a difficult position. She is rebellious	8		723	4750%		050%				050%		7396013_101.txt	61%
11	scene touches the speaker deeply and has a spiritual impact. 'The very houses	7		417	2375%		029%				029%		7421085_101.txt	29%
12	is uneducated but find out later that she has a master's in accounting. Yet she	8		461	1833%		020%				020%		6242929_101.txt	20%
13	Bev's courage and love of animals and has a respect for her and her works.	7		1,347	9053%		037%				037%		5751609_102.txt	87%
14	is a downtrodden, suppressed wife; she has a lazy husband, Jeremiah, who is	8		757	3056%		051%				051%		6915281_101.txt	61%
15	the speaker is in awe of the city and has a positive attitude towards it. He	8		890	4174%		038%				038%		6306862_101.txt	88%
16	gets Babamukuru to find her a job. She has a baby, gets a job and gets an	8		891	6825%		078%				078%		4835056_101.txt	78%
17	educated than her uncle and that she has a Master's degree and the money	8		851	3336%		076%				076%		4089007_101.txt	76%
18	the end of the novel when Nyasha has a psychotic episode she begins to	8		624	3120%		053%				053%		4189915_101.txt	54%
19	('gauntly'). He is unhealthy looking. He has a low self-esteem and does not	8		40	523%		0 3%				0 3%		4835056_101.txt	4%
20	spirited and whatever he wants, he has a way of getting. From the	8		196	1173%		017%				017%		7044322_101.txt	17%
21	by black males. His daughter has a voice being a member of white	8		1,744	65 8%		077%				077%		2819358_102.txt	77%
22	knows that his impoverished family only has a chance if he performs his duties	8		598	3753%		049%				049%		7396013_101.txt	50%
23	3 indicates that the subject is kingly or has a kingly characteristic or presence.	8		82	475%		0 7%				0 7%		6881395_101.txt	7%
24	and the prosecutor is thrilled as he has a tight case. Blood tests, the	8		1,568	8936%		075%				075%		8412618_101.txt	74%
25	The tone of the speaker has a sense of wishing and longing that	8		6	027%		0 1%				0 1%		3957630_102.txt	1%
26	soft to be capable of murder. She even has a soft spot for Mr Jay Gatsby. If	8		336	2136%		048%				048%		5761256_102.txt	50%
27	'looking down at them from behind', this has a figurative meaning to his	5		220	1350%		016%				016%		5780315_101.txt	17%
28	humbled by the city. He says the city has a 'mighty heart', which again	8		1,292	7532%		093%				093%		5780315_101.txt	94%
29	a capital letter for city as if the city has a name, like a person. He creates	8		998	5530%		072%				072%		5780315_101.txt	73%
30	. The title The Madonna of Excelsior has a connotation of the women. In the	5		373	2052%		053%				053%		6458182_101.txt	53%
31	'mighty heart'. It depicts that the city has a heart-beat, and it's powerful,	8		1,222	6947%		038%				038%		5780315_101.txt	89%
32	are still asleep it seems as if the city has a heart and this is now 'lying still'.	8		441	1791%		036%				036%		6915281_101.txt	36%
33	, and a cadaverous face, indispensable, has a dodgy look (gauntly unhealthy).	8		914	3959%		370%				036%		2014689_101.txt	86%
34	have been kept secret for so long? He has a newfound respect and	7		671	3214%		054%				054%		5614730_101.txt	54%
35	into the role of a subservient wife. She has a lot to say but cannot find the	8		945	28 7%		058%				058%		6288724_101.txt	58%
36	still has to do these chores – she has a lower rank than Babamukuru's	8		1,009	3834%		031%				031%		6915281_101.txt	81%
37	in this passage is Trotwood. Agnes has a very gentle character and feels	8		12	1 5%		0 1%				0 1%		6139621_101.txt	2%

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes

HAVE was coded as follows (see numbers in 'set'column):

1 auxiliary

2 operator

5 pseudo delexical

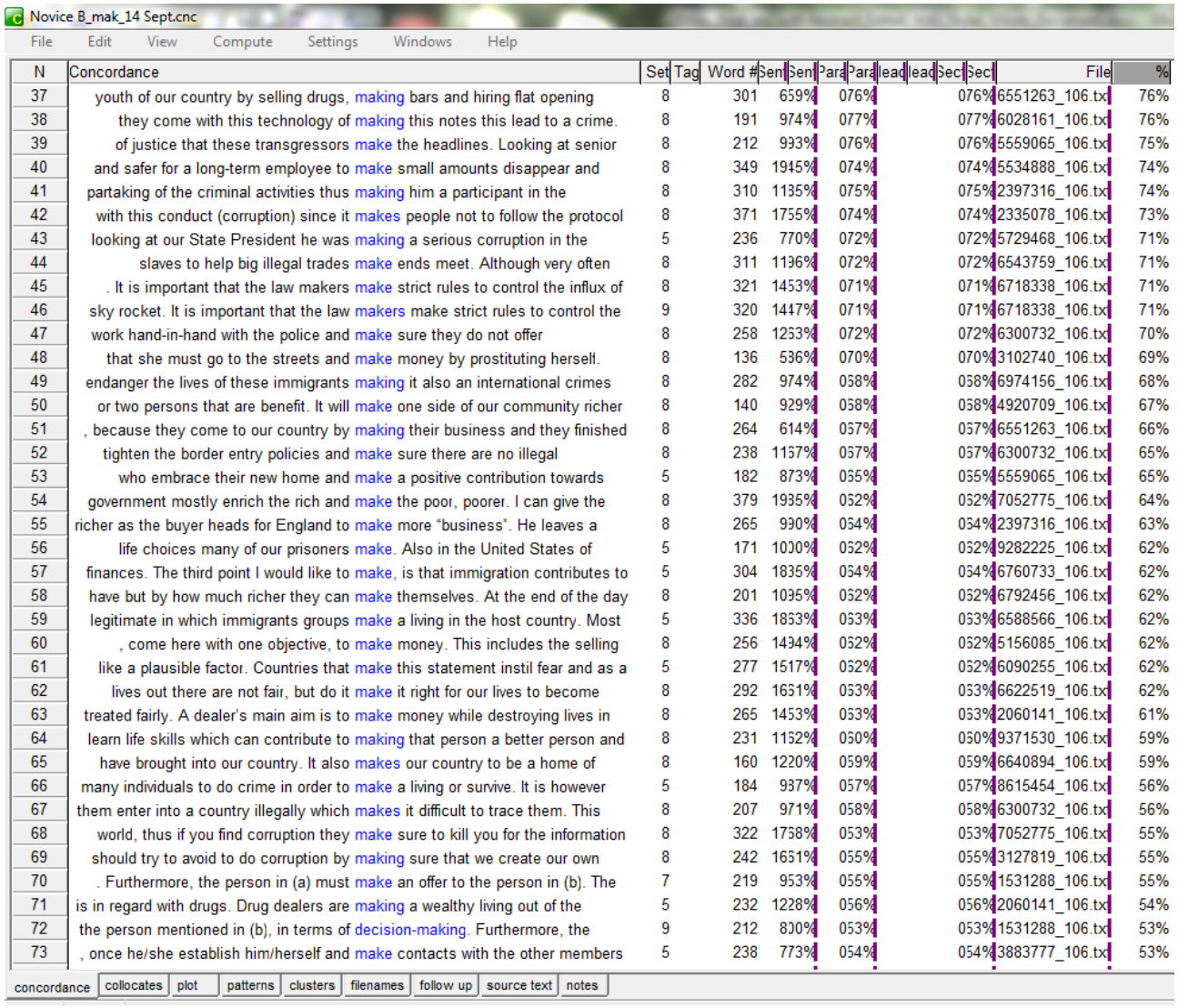
6 semi-modal

7 core delexical

8 lexical

9 anomalies (non-verbal)

Student Law *mak



Novice B_mak_14 Sept.cnc

File Edit View Compute Settings Windows Help

N	Concordance	Set	Tag	Word #	Sent	Sent	Para	Para	lead	lead	Sect	Sect	File	%
37	youth of our country by selling drugs, making bars and hiring flat opening	8		301	659%		076%				076%		6551263_106.tx	76%
38	they come with this technology of making this notes this lead to a crime.	8		191	974%		077%				077%		6028161_106.tx	76%
39	of justice that these transgressors make the headlines. Looking at senior	8		212	933%		076%				076%		5559065_106.tx	75%
40	and safer for a long-term employee to make small amounts disappear and	8		349	1945%		074%				074%		5534888_106.tx	74%
41	partaking of the criminal activities thus making him a participant in the	8		310	1135%		075%				075%		2397316_106.tx	74%
42	with this conduct (corruption) since it makes people not to follow the protocol	8		371	1755%		074%				074%		2335078_106.tx	73%
43	looking at our State President he was making a serious corruption in the	5		236	770%		072%				072%		5729468_106.tx	71%
44	slaves to help big illegal trades make ends meet. Although very often	8		311	1196%		072%				072%		6543759_106.tx	71%
45	. It is important that the law makers make strict rules to control the influx of	8		321	1453%		071%				071%		6718338_106.tx	71%
46	sky rocket. It is important that the law makers make strict rules to control the	9		320	1447%		071%				071%		6718338_106.tx	71%
47	work hand-in-hand with the police and make sure they do not offer	8		258	1253%		072%				072%		6300732_106.tx	70%
48	that she must go to the streets and make money by prostituting herself.	8		136	536%		070%				070%		3102740_106.tx	69%
49	endanger the lives of these immigrants making it also an international crimes	8		282	974%		058%				058%		6974156_106.tx	68%
50	or two persons that are benefit. It will make one side of our community richer	8		140	929%		058%				058%		4920709_106.tx	67%
51	, because they come to our country by making their business and they finished	8		264	614%		057%				057%		6551263_106.tx	66%
52	tighten the border entry policies and make sure there are no illegal	8		238	1137%		057%				057%		6300732_106.tx	65%
53	who embrace their new home and make a positive contribution towards	5		182	873%		055%				055%		5559065_106.tx	65%
54	government mostly enrich the rich and make the poor, poorer. I can give the	8		379	1935%		052%				052%		7052775_106.tx	64%
55	richer as the buyer heads for England to make more "business". He leaves a	8		265	990%		054%				054%		2397316_106.tx	63%
56	life choices many of our prisoners make. Also in the United States of	5		171	1000%		052%				052%		9282225_106.tx	62%
57	finances. The third point I would like to make, is that immigration contributes to	5		304	1835%		054%				054%		6760733_106.tx	62%
58	have but by how much richer they can make themselves. At the end of the day	8		201	1095%		052%				052%		6792456_106.tx	62%
59	legitimate in which immigrants groups make a living in the host country. Most	5		336	1853%		053%				053%		6588566_106.tx	62%
60	, come here with one objective, to make money. This includes the selling	8		256	1434%		052%				052%		5156085_106.tx	62%
61	like a plausible factor. Countries that make this statement instil fear and as a	5		277	1517%		052%				052%		6090255_106.tx	62%
62	lives out there are not fair, but do it make it right for our lives to become	8		292	1651%		053%				053%		6622519_106.tx	62%
63	treated fairly. A dealer's main aim is to make money while destroying lives in	8		265	1453%		053%				053%		2060141_106.tx	61%
64	learn life skills which can contribute to making that person a better person and	8		231	1152%		050%				050%		9371530_106.tx	59%
65	have brought into our country. It also makes our country to be a home of	8		160	1220%		059%				059%		6640894_106.tx	59%
66	many individuals to do crime in order to make a living or survive. It is however	5		184	937%		057%				057%		8615454_106.tx	56%
67	them enter into a country illegally which makes it difficult to trace them. This	8		207	971%		058%				058%		6300732_106.tx	56%
68	world, thus if you find corruption they make sure to kill you for the information	8		322	1758%		053%				053%		7052775_106.tx	55%
69	should try to avoid to do corruption by making sure that we create our own	8		242	1651%		055%				055%		3127819_106.tx	55%
70	. Furthermore, the person in (a) must make an offer to the person in (b). The	7		219	953%		055%				055%		1531288_106.tx	55%
71	is in regard with drugs. Drug dealers are making a wealthy living out of the	5		232	1228%		056%				056%		2060141_106.tx	54%
72	the person mentioned in (b), in terms of decision-making. Furthermore, the	9		212	800%		053%				053%		1531288_106.tx	53%
73	, once he/she establish him/herself and make contacts with the other members	5		238	773%		054%				054%		3883777_106.tx	53%

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes

MAKE was coded as follows (see numbers in 'set'column):

2 phrasal prepositional

4 prepositional

5 pseudo delexical

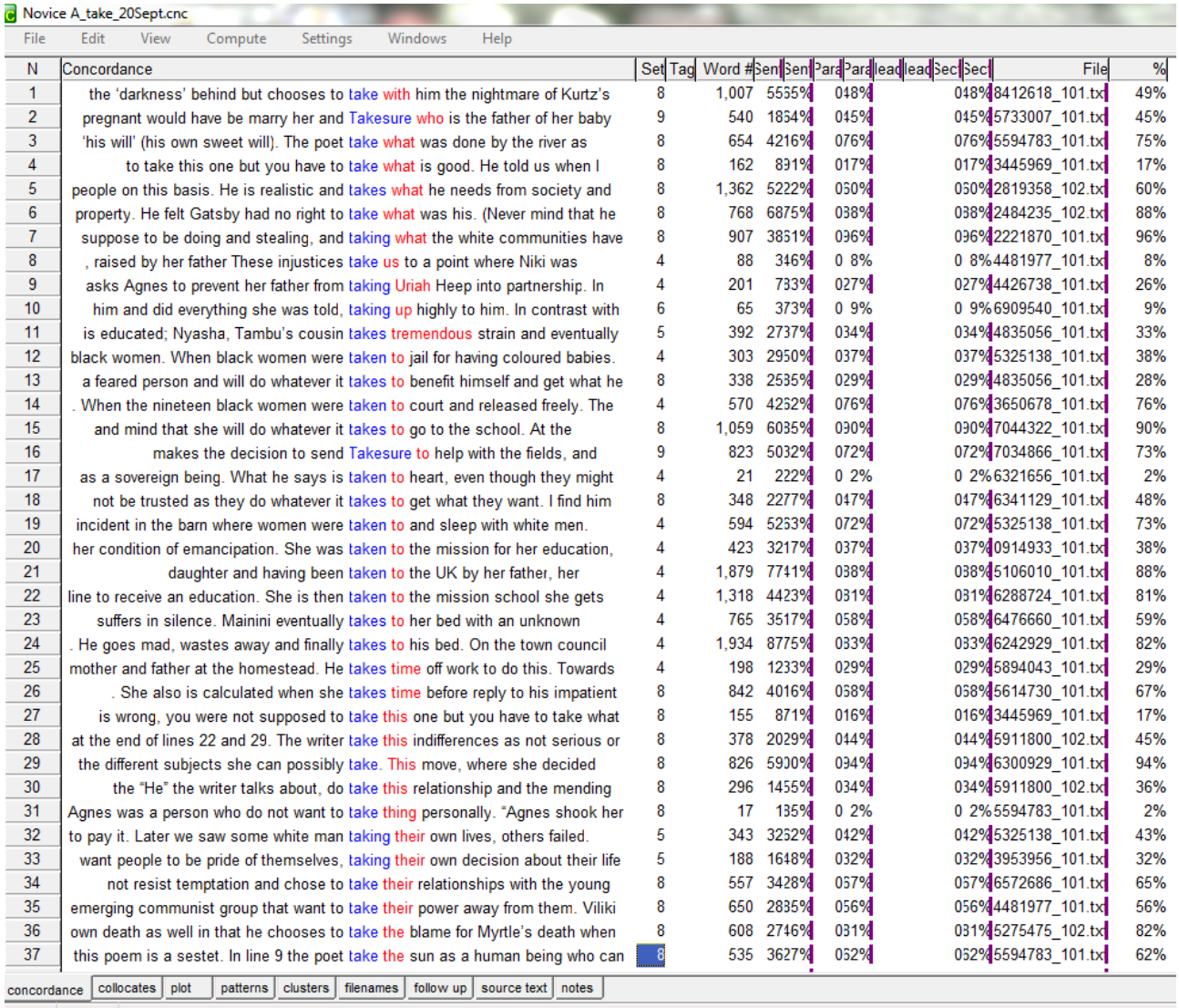
6 phrasal

7 core delexical

8 lexical

9 anomalies (non-verbal)

Student Lit *tak



Novice A_take_20Sept.cnc

File Edit View Compute Settings Windows Help

N	Concordance	Set	Tag	Word #	Sent	Sent	Para	Para	lead	lead	Sect	Sect	File	%
1	the 'darkness' behind but chooses to take with him the nightmare of Kurtz's	8		1,007	5555%		048%				048%		8412618_101.txt	49%
2	pregnant would have be marry her and Takesure who is the father of her baby	9		540	1854%		045%				045%		5733007_101.txt	45%
3	'his will' (his own sweet will). The poet take what was done by the river as	8		654	4216%		076%				076%		5594783_101.txt	75%
4	to take this one but you have to take what is good. He told us when I	8		162	891%		017%				017%		3445969_101.txt	17%
5	people on this basis. He is realistic and takes what he needs from society and	8		1,362	5222%		050%				050%		2819358_102.txt	60%
6	property. He felt Gatsby had no right to take what was his. (Never mind that he	8		768	6875%		038%				038%		2484235_102.txt	88%
7	suppose to be doing and stealing, and taking what the white communities have	8		907	3851%		096%				096%		2221870_101.txt	96%
8	, raised by her father These injustices take us to a point where Niki was	4		88	346%		0 8%				0 8%		4481977_101.txt	8%
9	asks Agnes to prevent her father from taking Uriah Heep into partnership. In	4		201	733%		027%				027%		4426738_101.txt	26%
10	him and did everything she was told, taking up highly to him. In contrast with	6		65	373%		0 9%				0 9%		6909540_101.txt	9%
11	is educated; Nyasha, Tambu's cousin takes tremendous strain and eventually	5		392	2737%		034%				034%		4835056_101.txt	33%
12	black women. When black women were taken to jail for having coloured babies.	4		303	2950%		037%				037%		5325138_101.txt	38%
13	a feared person and will do whatever it takes to benefit himself and get what he	8		338	2535%		029%				029%		4835056_101.txt	28%
14	. When the nineteen black women were taken to court and released freely. The	4		570	4252%		076%				076%		3650678_101.txt	76%
15	and mind that she will do whatever it takes to go to the school. At the	8		1,059	6035%		090%				090%		7044322_101.txt	90%
16	makes the decision to send Takesure to help with the fields, and	9		823	5032%		072%				072%		7034866_101.txt	73%
17	as a sovereign being. What he says is taken to heart, even though they might	4		21	222%		0 2%				0 2%		6321656_101.txt	2%
18	not be trusted as they do whatever it takes to get what they want. I find him	8		348	2277%		047%				047%		6341129_101.txt	48%
19	incident in the barn where women were taken to and sleep with white men.	4		594	5253%		072%				072%		5325138_101.txt	73%
20	her condition of emancipation. She was taken to the mission for her education,	4		423	3217%		037%				037%		0914933_101.txt	38%
21	daughter and having been taken to the UK by her father, her	4		1,879	7741%		038%				038%		5106010_101.txt	88%
22	line to receive an education. She is then taken to the mission school she gets	4		1,318	4423%		031%				031%		6288724_101.txt	81%
23	suffers in silence. Mainini eventually takes to her bed with an unknown	4		765	3517%		058%				058%		6476660_101.txt	59%
24	. He goes mad, wastes away and finally takes to his bed. On the town council	4		1,934	8775%		033%				033%		6242929_101.txt	82%
25	mother and father at the homestead. He takes time off work to do this. Towards	4		198	1233%		029%				029%		5894043_101.txt	29%
26	. She also is calculated when she takes time before reply to his impatient	8		842	4016%		058%				058%		5614730_101.txt	67%
27	is wrong, you were not supposed to take this one but you have to take what	8		155	871%		016%				016%		3445969_101.txt	17%
28	at the end of lines 22 and 29. The writer take this indifferences as not serious or	8		378	2029%		044%				044%		5911800_102.txt	45%
29	the different subjects she can possibly take. This move, where she decided	8		826	5900%		094%				094%		6300929_101.txt	94%
30	the "He" the writer talks about, do take this relationship and the mending	8		296	1455%		034%				034%		5911800_102.txt	36%
31	Agnes was a person who do not want to take thing personally. "Agnes shook her	8		17	135%		0 2%				0 2%		5594783_101.txt	2%
32	to pay it. Later we saw some white man taking their own lives, others failed.	5		343	3252%		042%				042%		5325138_101.txt	43%
33	want people to be pride of themselves, taking their own decision about their life	5		188	1648%		032%				032%		3953956_101.txt	32%
34	not resist temptation and chose to take their relationships with the young	8		557	3428%		057%				057%		6572686_101.txt	65%
35	emerging communist group that want to take their power away from them. Viliki	8		650	2835%		056%				056%		4481977_101.txt	56%
36	own death as well in that he chooses to take the blame for Myrtle's death when	8		608	2746%		031%				031%		5275475_102.txt	82%
37	this poem is a sestet. In line 9 the poet take the sun as a human being who can	8		535	3627%		052%				052%		5594783_101.txt	62%

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes

TAKE was coded as follows (see numbers in 'set'column):

2 phrasal prepositional

4 prepositional

5 pseudo delexical

6 phrasal

7 core delexical

8 lexical

9 anomalies (non-verbal)

APPENDIX F

DEVIANT MWUS AND ERRORS

MWUs are arranged according to degree of acceptability, from most acceptable to least acceptable

HAVE

STUDENT LIT

	Deviant MWU	Explanation of errors	Coding of errors	Acceptability
1	oom. He has changed because he **has sympathy** to his daughter.	Preposition after noun is incorrect. *To* should be *for*.	P	?
2	title The Madonna of Excelsior **has a connotation** of the women. [Context - In the novel, one may find a character who dominates the first part of the novel.]	D - should be definite article : *the connotation* because the context indicates that the definite article in *the woman* is incorrect as this is not meant to refer to a specific woman but to women in general.	D	?
3	visits her mother's shack. She **has plans** of taking her mother	P - preposition after noun incorrect. Should be *to* [to take].	P	?
4	but he does not. He goes on and **have sex** with Bev Shaw and pi	Concord error – should be *has* sex.	V – concord	?
5	which shows that he like to **have separation** between him a	SVC should be replaced by single-word V in passive voice, *be separated from*..	SVC	?
6	orced twice. He is in love and **have sex** with the minor students	V error – subject-verb agreement. *Have* should be *has*.	V – concord	?
7	body else. **The opinion that I have** on Uriah Heep is based	Preposition error – preposition *on* after noun should be *of.*	P	?
8	be laid upon him. The first one **have sex** with his student, but	V – subject-verb agreement error- *have* should be *has.*	V – concord	?
9	else's wife at the clinic and **have sex** with her. When he is	V –concord error: context revealed the subject of *have* to be *he* and thus *have* should be *has.*	V – concord	?
10	the theatre in the street. And **have sex** with her. The girl w	V – concord error: context revealed the subject of *have* to be *he* and thus *have* should be *has.*	V – concord	?
11	divorced, visits prostitutes and **have sex** with his student, yo	V error – concord error: *have* should be *has.*	V – concord	?
12	and pick up a street walker and **have sex** with her. I think	V – concord error: context revealed the subject of *have* to be *he* and thus *have* should be *has.*	V – concord	?
13	divorced twice, hired prostitute, **have sex** with the young student	V – concord error: context revealed the subject of *have* to be *he* and thus *have* should be *has.*	V – concord	?

14	patriarchal nature they did not **have choice** to say no to the	D - determiner error - missing indefinite article.	D – article	?
15	this fellow Uriah. He does not **have any trust** towards this m	P - –preposition *towards* after noun should be *in*. SVC – single-verb would have been preferable – *does not trust.*	P SVC	?
16	new subject which he does not **have any interest** on it. He i	Preposition error – preposition *on* after noun should be *in*.	P	?
17	t is lying still!' The poet is **having a hope** about the place	V - Incorrect use of progressive aspect present tense. P – preposition *about* after *hope* should be *for*.	V – tense P	?
18	and the people was dying. He **have no hope** of the universe	V – concord: subject-verb agreement *have* should be *has*. Preposition – preposition *of* after *hope* should be *for* or *in*.	V – concord P	?
19	to the idea. Though she still **had the respect** for what Baba	D – incorrect article: no article required.	D – article	?
20	assertive and thoughtful person who **has logical thinking** when awkward	V – collocation error. This is an example where a single-word verb would have been more appropriate - . ..*thinks logically*.	V – coll SVC	(*)
21	tive if someone ignore him. He **has an aggressive behaviour** while he talk.	V - collocation error – should be *shows aggressive behaviour*.SVC – a single-word verb with adverb would be more appropriate- *he behaves aggressively* D – incorrect use of article an with behaviour, an uncountable noun in this sense.	V – coll SVC D	(*)
22	They make the city of London **has life** like a living being.	D - indefinite article missing – *has **a** life.* V - construction *is alive* … would be more appropriate.	D – article V – SVC	(*)
23	d other question to Uriah. She **has full of respect**, she is n	*Has* used instead of *was/ or* she has a respect for. Context suggests student meant *is full of respect* and has used *has* incorrectly as past tense form of *is*. This is coded as a tense error.	V – tense	(*)
24	initely be why they mostly all **have a dependency** on him to work	Dependency is an UN (cannot take an article). This expression appears 21 times in the BNC, but in this context a single-word verb would be preferable, i.e. *depend*.	SVC	(*)
25	estead, although she was still **having feelings** with her family	V–tense: incorrect use of progressive aspect present tense; preposition after noun should be *for* or *about,* not *with*.	V – tense P	(*)
26	neighbours. The neighbours are **having the differences** between	Determiner error – inappropriate definite article. V error – incorrect progressive aspect present tense.	D – article V – tense	(*)

27	or hill, were untouched. He is **having the peaceful feeling** a	D - determiner error - definite article *the* should be indefinite *a.* V error – progressive aspect pres. tense instead of simple present. .	D – article V – tense	(*)
28	that David loves the women or **have love** of women. This is b	D - missing (indefinite) article; V – concord error: *have* should be *has* Preposition after noun - *of* should be *for.*	D – article V – concord P	(*)
29	them to do as he like. He do not **have a respect**, he is rude. N	D – inappropriate article: no article OR quantifier *any*.	D – article	(*)
30	Babamukuru's approval, she totally **had her dependence**. Even when	ADJ – adverb should be adjective *total*. SVC – single-word verb *she depended on* preferable.	ADJ SVC	(*)
31	to go to school. Tambu said she **had logical sense** than Nyasha	D - Missing quantifier *more*.	D – quantifier	(*)
32	Melanie's father. Lurie says he **have the disgrace** for his who	Verb - collocational error – disgrace collocates with *feel* or *experience* . V error – concord – *have* should be *has*. Determiner error - inappropriate definite article. SVC – single-word verb in passive voice – *was disgraced* - would be more appropriate.	V – coll D – article V – concord SVC	*
33	lines sestet where the poet **is having a solution** of his problem	V error- progressive aspect pres. tense instead of simple present. V - collocation error: solution collocates with *find*, not *have*. Preposition error: *of* after noun should be *to.*	V – tense V – coll P	*

STUDENT LAW

	Deviant MWU	**Explanation of errors**	**Coding of Errors**	**Acceptability**
1	Immigration **has had a major effect** to the	Preposition – *to* after noun *effect* should be *on.*	P	?
2	Lack of education therefore **has a direct correlation** to the	P – preposition *to* after noun *correlation* should be *with.*	P	?
3	by rich only and if you do not **have connection**, no need to g	D - indefinite article missing after *have*.	D – article	?
4	ration and the **implications it have** on the crime level of the	V - concord error – *have* should be *has*.	V – concord	?
5	Immigration **has had a major effect** to the rise	P – preposition after *effect* should be *on*, not *to*.	P	?

6	ing Live show, when Vuyo Mbuli **has the discussion** with John	D -Inappropriate definite article.	D – article	(*)
7	today's newspaper. Immigration **has had a great contribution**	V – collocation error: *contribution* collocates with *make*, not *have*. V deviation – *have* should be *make.*	V – coll	(*)
8	ctices. It is the rich who **are having these accesses** in most	V– inappropriate use of continuous aspect pres tense. Should be simple present. N - *access* is non-count noun (demonstrative error not counted as agrees with inmcorrect N).	V – tense N – number	(*)
9	South Africa, as a country **is having its own challenges** dealing	V - inappropriate use of progressive aspect pres. tense. Should be *has*.	V – tense	(*)
10	ding outside that was rich and **have influence** when it come to	V - concord error: *have* should be *has*.	V – concord	(*)
11	I would expect these people **to have a full disclosure** of the	V - collocation error: – disclosure collocates with *make*, not *have*.	V – coll	(*)
12	newspaper. Immigration **has had a great contribution** to	V – collocation error. Contribution collocates with *make*, not *have*.	V – coll	(*)
13	problem relating to crime **has had negative implications** on	P - prepositional error- should be *implications for*.	P	(*)
14	awful. Yes I agree some of them **have connection** with our guys	D – missing indefinite article after have, or N – *connection* should be plural.	D – article **OR** N	*
15	also been arrests of culprits **have deal** in drug trafficking	D– pronoun [*who*] missing. N – number: *deal* should be plural *deals.*	D – pronoun N	*

MAKE

STUDENT LIT

	Deviant MWU	**Explanation of errors**	**Coding of Errors**	**Acceptability**
1	tried to escape with her, he **made a huge mistake** of driving	D – incorrect indefinite article: should be *the*.	D – article	?
2	got in this world. David **made a lot of mistake** in life	N – should be plural to agree with '*a lot of*'.	N	?
3	The writer has reasoned and **makes a logical conclusion** th	V- collocation error: *conclusion* collocates with *reaches*, *arrives at*, not *makes*.	V – coll	(*)
4	ur with different eyes thus he **makes a conclusion** that will	V- collocation error: *conclusion* collocates with *reaches*, *arrives at*, not *makes*.	V – coll	(*)
5	she grows up. She is able **to make comparisons** of where she	P - preposition after *comparisons* should be *with*, *to* or *between.*	P	(*)
6	saw fit that **other decisions he make** are out of any sense, an	V – concord: *make* should be *makes.* .	V – concord	(*)
7	g the beauty of the City. This **makes a contrasts** with its use	N – should be singular after *a*.	N	(*)

8	t and fears him, **a realisation made** by the speaker later on	V-collocation error. *Realisation* collocates with *come to* or *reach*.	V – coll	(*)
9	h and manipulative because she **made conclusions** about Uriah	V - collocation error –*in this context conclusions* correlates with *jump to*.	V – coll	(*)
10	alerting Agnes to stand and **make voice be heard**. The narr	D – possessive pronoun missing.	D – pronoun	*
11	a councillor meanwhile she is **making his way** to the top by	D – possessive pronoun should be *her* in this context; V – inappropriate use of progressive aspect pres. tense in this context. Should be present tense.	D – pronoun . V – tense	*
12	t inquiry. But Lurie denied **to make statement** in writing. Th	D – missing indefinite article.	D – article	*

STUDENT LAW

	Deviant MWU	**Explanation of errors**	**Coding of Errors**	**Acceptability**
1	when a business manager or owner **make the wrong choice**. They e	V – concord error: *make* should be *makes*.	V – concord	?
2	ntion that their are coming **to make better living**. But not a	D – missing indefinite article.	D – article	?
3	rate due to the **movement they make** to other places because	SVC – single-word V such as *migrate* would have been more appropriate, or even a noun such as *migration*.	SVC	?
4	d with drugs. Drug dealers are **making a wealthy living** out of	Adj – collocation error: *living* collocates with *good*, *successful,* not *wealthy* in this context.	ADJ – coll	?
5	/she establish him/herself and **make contacts** with the other	N – *contacts* should be singular.	N – number	(*)
6	ideas. Some they are coming **to make living**, with some stealing	D – missing indefinite article.	D – article	(*)
7	ion where they commit crime **to make living** in foreign country	D – missing indefinite article.	D – missing article	(*)
8	ls in this day and age. Try **to make difference**, do not just	D- missing indefinite article;.	D – article	(*)
9	l for **the corruption that they made**. If some of us are aware	V - collocation error: *corruption* collocates with *commit,* not *make*.	V – coll	(*)
10	seated in these positions and **made corruption** is how they g	V - collocation error: *corruption* collocates with *commit,* not *make*.	V – coll	(*)
11	the pocket. **Corruption is only making** by those who are employ	V –tense: *is making* should be *made*. V-collocation error. Corruption collocates with *commit*, not *make*.	V – tense V – coll	*
12	former president of Justice was **making a serious corruption**,	V – tense error: inappropriate use of progressive aspect perfect tense. D – inappropriate article: *corruption* UC. V collocation error: *corruption* collocates with *commit* not *make*.	V – tense D – article V – coll	*
13	ion. They support the BEE's in **making corruption** for their o	V – collocation error: *corruption* collocates with *commit*, not *make*.	V – coll	*

14	s not rich he or she could not **make corruption** she or he has	V – collocation error: *corruption* collocates with *commit*, not *make*.	V – coll	*
15	at our State President he **was making a serious corruption** i	V - collocation error: *corruption* collocates with *commit*. D-inappropriate article; V - tense error: inappropriate use of progressive aspect perfect tense. Context suggests it should be simple past tense.	V – coll D – article V – tense	*
16	are no **new development that have been made** there for years ago until	V –concord error OR N – should be plural.	V – concord OR N – number	*
17	they come with **this technology of making** this notes this lead to	D – inappropriate demonstrative this.	D – demonstrative	*

TAKE

STUDENT LIT

	Deviant MWU	**Explanation of errors**	**Coding of Errors**	**Acceptability**
1	at she and her family would be **taken care** off if she does no	P – preposition *off* after *care* should be *of*. [Spelling error]	P	?
2	eir places. He was the one who **takes decisions** on how the ho	V – tense: *takes* should be *took*.	V – tense	?
3	ple to be pride of themselves, **taking their own decision** abo	N – number: *decision* should be plural.	N	?
4	weaknesses, fostered them, and **take advantage** of them, until	*V* – concord Error. Context indicates *take* should be *takes*.	V – concord	(*)
5	ment. Agnes claims that he has **taken advantages** of her father	Noun – should be singular: a*dvantage* is UC in this context.	N	(*)
6	themes of Lucia. Babamukuru was **taking care** of Maiguru, he tr	V – tense: context indicates inappropriate use of progressive aspect. Should be perfect tense.	V – tense	(*)
7	le's input in the **decisions he took** with their lives. She ki	P – preposition incorrect. Could be replaced by *in* or *regarding their lives.*	P	(*)
8	er Myrtle body. Wilson without **taking much longer time** he went	ADJ – incorrect adjective: *much longer* should be *more*. SVC – single-word verb more appropriate such as *hesitating*.	ADJ SVC	*
9	because she has no qualms for him **taking the blame** of running Myrtle	P – Preposition of after blame should be *for*.	P	*
10	good and in deep humility. He **takes conversation** with other	V - collocation error – *conversation* collocates with *make*, not *take.*	V – coll	*
11	away, advice that he does not **take hid** of. Gatsby could als	V – error in verb: *hid* should be *heed*.	V – word	*
12	l says that it has no power to **take and decision** but after f	D- indefinite article missing; incorrect use of *and*.	D – article	*

STUDENT LAW

	Deviant MWU	**Explanation of errors**	**Coding of Errors**	**Acceptability**
1	at all but let justice and law **takes its effect.**	D – possessive pronoun should be plural, *their*. Verb – mood: verb is subjunctive because of *let*, which requires a non-finite form of the verb, the infinitive, *take*	D – pronoun V – mood	?
2	lack of corrective **action that take place**. Corrective procedures	Verb – concord error: should be *takes*.	V – concord	?
3	in many cases **actions are not taken** against the person involved	N – should be singular: *take action*	N-number	?
4	ny. And the government was not **taking any action** on that case	V - tense. Context indicates inappropriate use of progressive aspect present tense. Should be perfect tense.	V - tense	(*)
5	ure that their government must **take actions** against them.	N – number: should be singular *action*	N – number	(*)
6	requires money and a person to **take risk**, but the only plan	N – should be plural *risks* OR D – missing indefinite article.	N – number **OR** D – article	(*)
7	business suits who choose to **take part in corruption**, who	V - collocation error: *corruption* collocates with *commit*, not *take part in.*	V – coll	(*)
8	they can get by promoting and **taking part in crimes**. Example	V - collocation error: *crimes* collocates with *commit*, not *take part in.*	V – coll	(*)
9	g here in South Africa, he was **take care** of neighbouring countries	Verb – tense/aspect error. Should be *taking care of*.	V – tense	(*)
10	any relevant **punishment must be taken** according and they must	V - collocation – in this context *punishment* does not collocate with *taken* but with *meted out*. ADJ – inappropriate adjective. Should be adverb *accordingly*.	V – coll ADJ	(*)
11	e rich people some of them **are taking an advantage** of poor p	V – tense: Inappropriate use of progressive aspect – should be *take advantage*. D – inappropriate indefinite article.	V – tense D – article	(*)
12	The rich **take advantage** of using their	P – preposition after *advantage* should be *by.*	P	(*)
13	greedy. Greed is from the devil; **take a look** of the following	P – preposition after look should be *at.*	P	(*)
14	do believe that if **corruption take a lead** it will takes us	V - collocation error: in this context *take* collocates with a *hold*; or replace with expression *gets the upper hand* V – concord error: *take* should be *takes*.	V – coll V – concord	(*)
15	and get hold of that. The state **took a further steps** against	D – inappropriate indefinite article. OR N – number: singular *step*.	D **OR** N	(*)

16	and very little serious **crimes takes place** like murder or rape	Verb – concord error: *very little* [should be *few*] *crimes* is plural, verb should be *take place.* SVC – single-word verb such as commit (passive) would be more appropriate.	V – concord SVC	*
17	nes of some Africans also **were taken rescued** by the Spanish	SVC – simple V *rescue* would be more appropriate here.	SVC	*
18	people they treat them badly and **taking an advantage** of them.	V - tense: should be *take*. D – inappropriate indefinite article.	V – tense D – article	*
19	the rich people because **poverty takes part**. In most cases of	V - collocation error: *poverty* collocates with *play*, not *take part in.* D – missing indefinite article.	V – coll D – article	*
20	o anything achieve that. If it **take to bribe** somebody in order	S – structural error: should read *If a bribe is what it takes?* or *If it takes a bribe (to somebody)?*	S	*

APPENDIX G

SAMPLE SPSS OUTPUT

```
SPLIT FILE OFF.
MEANS TABLES=TotalPerc BY Genre BY Gender
  /CELLS=MEAN COUNT STDDEV.
```

Means

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Total % * Course type * Gender	298	100.0%	0	0.0%	298	100.0%

Report

Total %

Course type	Gender	Mean	N	Std. Deviation
Literature	Male	64.12	28	19.326
	Female	69.20	111	17.694
	Total	68.18	139	18.078
Law	Male	53.67	95	17.272
	Female	66.29	64	17.909
	Total	58.75	159	18.544
Total	Male	56.05	123	18.217
	Female	68.14	175	17.777
	Total	63.15	298	18.894

```
GET
  FILE='C:\Users\Scheera\Documents\Doctorate\D Litt et Phil\2012\Analysis Jun
e 2012\SPSS\Revised data_14June2012.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
T-TEST GROUPS=Gender(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=TotalPerc
  /CRITERIA=CI(.95).
```

T-Test

```
[DataSet1] C:\Users\Scheera\Documents\Doctorate\D Litt et Phil\2012\Analysis
June 2012\SPSS\Revised data_14June2012.sav
```

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Total %	Male	123	56.05	18.217	1.643
	Female	175	68.14	17.777	1.344

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
Total %	Equal variances assumed	.027	.871	-5.720	296
	Equal variances not assumed			-5.695	258.709

Independent Samples Test

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
Total %	Equal variances assumed	.000	-12.087	2.113
	Equal variances not assumed	.000	-12.087	2.122

Independent Samples Test

		t-test for Equality of Means	
		95% Confidence Interval of the Difference	
		Lower	Upper
Total %	Equal variances assumed	-16.246	-7.928
	Equal variances not assumed	-16.266	-7.908

```
SORT CASES  BY Genre.
SPLIT FILE SEPARATE BY Genre.
T-TEST GROUPS=Gender(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=TotalPerc
  /CRITERIA=CI(.95).
```

T-Test

Course type = Literature

Group Statistics[a]

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Total %	Male	28	64.12	19.326	3.652
	Female	111	69.20	17.694	1.679

a. Course type = Literature

Independent Samples Test[a]

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
Total %	Equal variances assumed	.182	.670	-1.333	137
	Equal variances not assumed			-1.264	39.195

Independent Samples Test[a]

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
Total %	Equal variances assumed	.185	-5.082	3.812
	Equal variances not assumed	.214	-5.082	4.020

Independent Samples Test[a]

		t-test for Equality of Means	
		95% Confidence Interval of the Difference	
		Lower	Upper
Total %	Equal variances assumed	-12.621	2.457
	Equal variances not assumed	-13.212	3.047

a. Course type = Literature

Course type = Law

Group Statistics[a]

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Total %	Male	95	53.67	17.272	1.772
	Female	64	66.29	17.909	2.239

a. Course type = Law

Independent Samples Test[a]

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
Total %	Equal variances assumed	.243	.623	-4.451	157
	Equal variances not assumed			-4.419	131.960