Predicting IELTS Ratings using Vocabulary Measures

Theodosia Demetriou

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy

Contents

ABSTRACT

This thesis addresses the relationship between vocabulary measures and IELTS ratings. The research questions focus on the relationship between measures of lexical richness and teacher ratings. The specific question the thesis seeks to address is: Which measures of lexical richness are the best for predicting the ratings? This question has been considered central in vocabulary measurement research for the last decades particularly in relation to IELTS, one of the most popular exams in the world. Therefore, if a model can predict IELTS scores by using vocabulary measures it could be used as a predictive tool by teachers and researchers worldwide.

The research was carried out through two studies, Study 1 and Study 2 and then the model was tested through a third smaller study. Study 1 was a small pilot study which looked at both oral and written data. Study 2 focused on written data only. Measures of both lexical diversity and sophistication were chosen for both studies. Both studies followed similar methodologies with the addition of an extra variable in the second study. For the first study data was collected from 42 IELTS learners whereas for the second study an existing corpus was used. The measures investigated in both studies were: Tokens, TTR, D, Guiraud, Types, Guiraud Advanced and P_Lex. The first four are measures of lexical diversity, the other three measures of lexical sophistication. However, all of the previous measures are measures of breadth of vocabulary. For the second study, a measure of formulaic count was added. This is an aspect of depth of vocabulary used to check if results would improve with this addition. Formulaic sequences were counted in each essay by using Martinez and Schmitt's (2012) PHRASE List of the 505 most frequent non-transparent multiword expressions in English.

The main findings show that all the measures correlate with the ratings but Tokens has the highest correlation of all lexical diversity measures, and Types has the highest correlation of all lexical sophistication measures. TTR, Guiraud and P_Lex can explain 52.8% of the variability in the Lexical ratings. In addition, holistic ratings can be predicted by the same two lexical diversity measures (TTR and Guiraud) but with a different measure of lexical sophistication, Guiraud Advanced. The model consisting of these three measures can explain 49.2% of the variability in the holistic

ratings. The formulaic count did not seem to improve the model's predictive validity, but further analysis from a qualitative angle seemed to explain this behaviour. In Study 3, the holistic ratings model was tested using a small sample of real IELTS data and the examiners comments' were used for a more qualitative analysis. This revealed that the model underestimated the scores since the range of ratings from the IELTS data was wider than the range of the data from Study 2 which were used as the basis for the model. This proved to be a major hindrance to the study. However, the qualitative analysis confirmed the argument that vocabulary accounts for a high percentage of variance in ratings and provided insights to other aspects that may influence raters which could be added to the model in future research. The issues and limitations of the study and the current findings contribute to the field by stimulating further research into producing a predictive tool that could inform students of their predicted rating before they decide to take the IELTS exam. This could have potential financial benefits for students.

ACKNOWLEDGEMENTS

GLOSSARY OF TERMS AND ABBREVIATIONS

Advanced TTR: The advanced TTR is a measure proposed by Daller, Van Hout and Treffers-Daller in 2003. This ratio is calculated by dividing the number of advanced types by the number of tokens.

AWL: Academic Word List (Coxhead, 2000).

BNC: British National Corpus.

Carroll's CTTR: The corrected TTR or CTTR was proposed by Carroll in 1964 and it is calculated by dividing the number of types by the square root of twice the number of tokens.

CHAT: Codes for Human Analysis of Transcripts (MacWhinney, B., 2000). This is the transcription and coding format used in CHILDES.

CHILDES: Child Language Data Exchange System. A database for sharing and researching conversational interactions.

CLAN: Stands for Computerised Language Analysis. This programme was developed by MacWhinney (2000) to analyse data in the format of CHILDES. It comprises of various commands for analysing language including *vocd*.

Coh-Metrix: A computational tool used to calculate the coherence of texts with a wide range of measures. It replaces common readability formulas by applying the latest in computational linguistics and linking this to the latest research in psycholinguistics (University of Memphis website).

Collocation: When a sequence of words co-occurs more often than would be expected by chance.

Compound Words: A compound word is formed when two words are joined to create a new word, for example: post office.

Concurrent Validity: 'A type of validity which is concerned with the relationship between what is measured by a test (usually a newly developed test) and another existing criterion measure, which may be a well-established test, a set of judgements or some other quantifiable variable' (Davies, Brown, Elder, Hill, Lumley and McNamara, 1999:30).

Content Words: Nouns, 'full' verbs, adjectives and adverbs.

Construct Validity: 'The construct validity of a language test is an indication of how representative it is of an underlying theory of language learning' (Davies et al., 1999:33).

Content Validity: This refers to the extent to which a test measures what it intended to.

Convergent Validity: 'A type of validity which is concerned with the similarity between two or more tests which are claimed to measure the same underlying trait or ability' (Davies et al, 1999:34).

Corpus/Corpora: A large set of texts that is usually stored electronically so that is easily analysable.

Corrected TTR: See Carroll's TTR.

Correlation: 'A procedure which measures the strength of the relationship between two (or more) sets of measures which are thought to be related' (Davies et al, 1999:35).

Cronbach's Alpha: A measure of internal consistency or reliability, which can take values between negative infinity and 1 (1=maximum .06 or .07 is often seen as the lower limit).

D: Developed by Malvern and Richards, D is a new measure of lexical diversity designed to overcome the sample size problem of TTR (See Malvern and Richards, 2002).

Descriptive Statistics: Summary data of the group measured.

Diagnostic Test: A diagnostic test identifies a learner's strengths and weaknesses. It is not used as much as other tests (that provide general information) because it is time-consuming and difficult to develop and administer.

EAP: English for Academic Purposes.

EFL: English as a Foreign Language.

Extrinsic Measures: Measures of lexical richness that look beyond just counting words and are based on the frequency/sophistication of a word. This term was used by Meara and Bell (2001).

Formulaic Language: The use of idioms, collocations, turns of phrase, routines, set phrases, rhymes, prayers and proverbs in speech. (Cardiff University Website, www.cardiff.ac.uk) 'Formulaic language is a term used by many researchers to refer to the large units of processing- that is, lexical units that are more than one word long' (Wray, 2008:3).

Formulaic Sequence: The generic term used to describe instances of formulaic language such as lexical bundles, phrasal expressions etc.

Function Words: Articles, prepositions, pronouns, conjunctions and auxiliaries. These have little if any meaning in isolation. They belong to the grammar of the language rather than its vocabulary (Read, 2000).

GSL: General Service List of English words. This consists of the two thousand more useful word families in English (West, 1953).

Guiraud's Index (Guiraud): In 1960, Guiraud proposed a measure which was an alteration of the TTR. It represents the number of types divided by the square root of the number of tokens (Guiraud, 1960).

Guiraud Advanced: Measure of lexical sophistication proposed by Daller, Van Hout and Treffers-Daller (2003). It is calculated by dividing the advanced types (words that are not in the basic lists, as defined by Nation) by the square root of the number of tokens (all tokens, not advanced tokens).

Halo Effect: 'The tendency of a rater to let an overall judgement of the person influence judgements on more specific attributes…For example, in speaking tests where raters are asked to assess a single performance according to a number of different criteria (e.g. accuracy, fluency, intelligibility, appropriateness) these ratings are often closely aligned' (Davies et al., 1999).

Heap's Law: Heap's Law (1978) describes the number of distinct words in a text as a function of the text's length.

Herdan's Index: Herdan's Index (1960), or LogTTR is calculated by dividing the logarithm of tokens by the logarithm of types.

Holistic Rating: Global rating: 'A type of marking procedure which is common in communicative language testing whereby raters judge a stretch of discourse (spoken or written) impressionistically according to its overall properties rather than providing separate scores for particular features of the language produced' (Davies et al, 1999:75).

Idiom: 'An idiom is an expression whose meaning cannot always be really derived from the usual meaning of its constituent elements. It is hard to tell from the literal meaning of the individual words, for example, that *to kick the bucket* or *to bite the dust* means to die' (Cooper, 1999:233).

IELTS: International English Language Testing System. A test designed to assess the language proficiency of non-native speakers of English who wish to enter English

tertiary education. It covers all four receptive and productive skills (listening, reading, writing, and speaking). The results are reported on a 9-point scale with nine being the highest mark that can be awarded.

Inflected Forms: Modified forms of words used to produce different grammatical categories, such as tense or plural form (for example eat- eat*s*, play- play*ed*).

Inferential Statistics: 'Methods used in making general probabilistic statements about the population under investigation on the basis of what is known about a sample of that population' (Davies et al., 1999:81).

Inter-rater Reliability: Shows the extent to which two or more raters' judgements agree (level of consensus) when rating learners' performance in tests.

Intrinsic Measures: This term was used by Meara and Bell (2001) to refer to measures of lexical richness that are based on tokens and types.

L1: First Language, also known as 'mother tongue' or native language.

L2: Second Language.

Lambda: Lambda values are produced in P_Lex (Meara and Bell, 2001). Lambda is a single parameter from a Poisson distribution. Poisson distribution is the probability of obtaining exactly *n* successes in *N* trials (e.g. 4 rare words in 10 words). Lambda values in P_Lex normally range from 0 to 4.5, and the higher the figure, the higher the proportion of infrequent words.

Lemma: The base and inflected forms of a word, for example: play, plays, played and playing (Read, 2000).

Lemmatisation: The process in which words are counted as lemmas (without all the inflected forms). Therefore, the words *play, played,* and *players* will be counted as one type (*play).*

Lex30: Lex30 is a word association task which stimulates vocabulary production. Word frequency data is used to measure the vocabulary produced. It was proposed by Meara and Fitzpatrick (2000). Lex30 is a test of productive vocabulary.

Lexeme: The base form of a word (as it is found in the dictionary).

Lexical Bundle: This term was introduced by Biber, Johansson, Leech, Conrad, and Finegan (1999) and refers to words that repeatedly occur together.

Lexical Diversity: Also known as lexical variation. These two terms are interchangeable. It refers to the amount of repetition in a text (it indirectly refers to vocabulary size).

Lexical Density: One of the dimensions of lexical richness proposed by Read (2000) which refers to the use of a higher percentage of content words rather than function words.

Lexical Richness: A term used by Read (2000) to describe the effective use of vocabulary in good writing. Lexical richness consists of four components: lexical variation, lexical sophistication, lexical density, and low number of errors. (Read, 2000). Malvern, Richards, Chipere and Duran (2004) use the term to describe someone's vocabulary in terms of lexical diversity (or lexical variation) and lexical sophistication (this is the term used in this study).

Lexical Sophistication: One of the dimensions of lexical richness (Read, 2000; Malvern et al., 2004) which refers to the use of 'rare' or infrequent words.

Lexical Variation: One of the dimensions of lexical richness (Read, 2000; Malvern et al., 2004). It refers to the number of different words used in a text (not repeated words).

LFP: Lexical Frequency Profile. Developed by Laufer and Nation (1995), LFP is a tool used for assessing the use of low frequency words by EFL learners, allocating

all the words of an essay into four different frequency bands. LFP gives the proportion of infrequent words in the text (Malvern et al, 2004:193).

Maas Index: Maas (1972) proposed this index of lexical diversity which is a logarithmic transformation of the type token ratio.

MELAB: Michigan English Language Assessment Battery.

MEU: Morpheme Equivalent Unit.

MSTTR: Mean Segmental Type- Token Ratio. A method proposed by Johnson (1944) to overcome the problem with sample size, that TTR seems to have (Malvern et al., 2004:196).

MWU: MultiWord Unit (Pawley and Syder, 1983).

N: The symbol used in formulas to refer to the number of tokens.

NDW: A simple measure of lexical variation. It represents the number of different words in a sample. This measure gives a single value (Malvern et al., 2004).

Operationalise: To operationalise means to turn the construct (the theoretical model of language proficiency underlying the test) into something that can be tested. 'Tests themselves can be viewed as operationalisations of the test construct' (Davies et al., 1999:136).

Outlier: This is a score that does not belong with the rest of the scores (an extreme score that does not fit with the general pattern of behaviour).

Pearson's *r*: This is a suitable correlation for dealing with interval type variables.

PHRASE List: The PHRASal Expressions List was developed by Martinez and Schmitt in 2012 and consists of the 505 most frequent non-transparent multiword expressions in English.

Phrasal Verb: A verb that consists of more than one word, such as *put up with.*

P_Lex: P_Lex is a computer programme that models the occurrence of rare words with a Poisson distribution. P_Lex was developed by Meara and Bell (2001) and is a measure of lexical sophistication.

Predictive Validity: 'Measures how well a test predicts performance on an external criterion' (Davies et al., 1999:149).

Predictor: A measure (often a test) that is used to predict if there is an effect on another variable.

Range: Software developed by Laufer and Nation (1995). It sorts the vocabulary of different texts into four different lists (frequency layers) and gives an LFP (Lexical Frequency Profile) which shows the richness of each text.

Rater: 'The judge or observer who operates a rating scale in the measurement of oral and written proficiency' (Davies et al., 1999:161).

Regression: 'A statistical technique which calculates the relationship between two or more variables and hence allows predictions to be made about performance on one variable on the basis of information about performance on another' (Davies et al, 1999:165).

Reliability (or test reliability): This refers to the agreement between the results of a test with itself or with another test. In order for a test to be reliable, the same results should be acquired when the test is repeated.

SLA: Second Language Acquisition.

Spearman's Rho: It is the non-parametric equivalent to the Pearson correlation.

SPSS: Statistical Package for Social Sciences. A statistical programme used for data analysis.

Stepwise Regression: 'A technique for performing a multiple regression analysis whereby variables are entered one by one, or step by step into the equation until the best model (i.e. the one in which the greatest proportion of variance is explained) is arrived at' (Davies et al 1999:189).

Tokens: The total number of words produced by someone in a piece of writing or oral speech.

TOEFL: Test of English as a Foreign Language. This is a test of language proficiency designed for second language learners that wish to attend American universities.

Types: The number of different words produced by someone in a piece of writing or oral speech.

TTR: Type Token Ratio. A measure of lexical diversity which is calculated by dividing the number of types by the number of tokens in the text.

UWL: University Word List (Xue and Nation, 1984).

V: The symbol used in formulas to refer to the number of types.

Validity (or test validity): A test is valid if it provides an accurate representation of an abstract concept such as proficiency (Davies et al., 1999).

VKS: Vocabulary Knowledge Scale. An instrument developed by Paribakht and Wesche (1993) which captures in an efficient way certain stages in the initial development of given words. This is a measure of depth of vocabulary knowledge.

Vocabulary Levels Test (VLT): This is a test of written receptive vocabulary size developed by Nation (1983; 1990).

VocabProfile: The computer programme used to produce LFPs.

Vocd: The command in CLAN used to compute D (the software was developed by McKee, Malvern and Richards (2000).

Word Associates Format (WAF): A test of depth of vocabulary knowledge developed by Read (1993).

Word Family: Words that share a common base. Different prefixes and suffixes can be added to this base to create new words. This group of words is called a Word Family.

CHAPTER 1- INTRODUCTORY CHAPTER

1.1 BACKGROUND TO THE THESIS

This thesis is an investigation of the extent to which IELTS ratings can be predicted by measuring richness of vocabulary. The thesis comprises three studies: Study 1, a pilot study based on original data: Study 2, a complete re-analysis of an existing corpus, and Study 3, which uses data from the IELTS website. My research develops previous work by Read and Nation (2002), one of the seminal studies of IELTS ratings and lexical richness, and aims to shed further light on the relationship between measures of vocabulary knowledge and IELTS ratings.

Recent decades have seen increased academic attention to the field of vocabulary and vocabulary measurement (for example Nation, 1983; Meara and Buxton, 1987; Schmitt, 1994; Daller, Milton and Treffers-Daller, 2007; Milton, 2009), and there have also been indications that vocabulary plays an important role in proficiency ratings, all of which interested me particularly, through my work as a linguist and English teacher. I therefore decided to investigate the relationship between measures of vocabulary and scores achieved in the IELTS exam, one of the most popular exams worldwide.

IELTS is one of the fastest growing tests in the world (McGovern and Walsh, 2006), with currently around a million candidates each year. 'The IELTS test… has experienced an increase in the number of test sittings from about 20,000 a year after its inception in 1989 to approximately 220,000 sittings in 2001' (UCLES, 2002a, in Moore and Morton, 2005:44). Every day, people from all over the world choose to take this particular test as it is one of the most widely accepted methods of assessing academic English skills, making it a very appropriate object for study. Additionally, Lexical Resource is one of the criteria used for rating IELTS writing and speaking responses.

Read and Nation (2002) analyse the lexical statistics of a corpus of IELTS Speaking tests, looking at the characteristics of vocabulary use at different score band levels, including the different kinds of formulaic language used, and whether this varied at

different band score levels. Despite some similarities in our studies, this thesis notably includes a different lexical richness measure and only uses written data in the main study (Study 2). This will later be discussed in detail in relation to the thesis findings.

Several lexical knowledge theories are discussed in the thesis but the theoretical issue of particular significance is that we can analyse vocabulary from two perspectives: breadth vs. depth and receptivity vs. productivity. Depth of vocabulary knowledge is approached in this study from a components perspective (the use of collocations/formulaic language is one of the components of depth of knowledge). Breadth of vocabulary knowledge in this study is defined as the size of vocabulary (determined both by lexical diversity and lexical sophistication measures). The starting point of the research is that vocabulary richness can be measured in different ways, but a combination of lexical diversity and lexical sophistication measurements gives a better overview than a single measurement. Vocabulary richness is closely and significantly associated with language proficiency and ratings (band scores), therefore in this study I have attempted to create a model based on various lexical richness measures to predict IELTS ratings. It needs to be clarified from the beginning that for the purposes of the present study lexical richness is used as a generic term to describe vocabulary knowledge (even though as explained in the thesis, lexical richness can also be a term referring to a characteristic of a text, while vocabulary knowledge is a general term to describe someone's knowledge), and lexical diversity and sophistication are two of the dimensions of lexical richness. Both these aspects belong to the breadth dimension of vocabulary knowledge. Furthermore, an aspect of depth of knowledge (use of formulaic sequences/collocations) was also added to the model in an attempt to increase the model's predictive validity. Formulaic language should hold a prominent place in vocabulary research according to Schmitt (2010:9), as both written and spoken discourse consist of large amounts of formulaic language (as large as 52-58% according to Erman and Warren, 2000). Regarding use of terminology, it is common in the literature for different terms to be used interchangeably (Fatahipour, 2012) in relation to formulaic language. Specifically, 'formulaic sequence' is used as a generic term that covers all different types of formulaic language. Collocations are an

example of formulaic language therefore references to formulaic sequences (or formulaic language) in this thesis are also indirect references to collocations.

One of the main studies regarding lexical richness measures and the use of formulaic sequences is Fatahipour's (2012), which investigates the possibility of a correlation between measures of lexical richness and language ability, and between measures of lexical richness and frequency of formulaic sequences in participants' essays, using the same phrase list (Martinez and Schmitt, 2012) as the present study.

Unexpected complications which are discussed in the work led to a change in the research design. Specifically, the main difficulty encountered was the fact that after the small pilot study (Study 1), in which IELTS essays from mock exams were collected and analysed, it was very important to use a larger sample of authentic IELTS data for Study 2. Given that Read and Nation had used real IELTS data in their study, it seemed likely that Cambridge would grant me access to essays and band scores from their databank, if they were used on an anonymous basis. However, it soon became clear that it would not be possible to gain access to their data. The response was very negative and IELTS seemed very secretive about the test and band scores. Hence I had to compromise with using other data. Turlik's (2008) corpus was the best alternative as the essays collected (even though they were not taken from actual IELTS exams) were marked by trained IELTS examiners and could help me build my predictive model. However, this corpus did not include a wide range of ratings (scores), and this seemed to influence the performance of the model. When the model was tested with 8 real IELTS essays, taken from the IELTS website and involving a wider range of ratings, it seemed to underestimate the values and was only successful in predicting lower-marked essays. These problems are discussed in detail in Chapter 7, however the process showed that a fully functional model can be created if real IELTS data are used. Despite this limitation, the study managed to show that vocabulary plays a major role in IELTS ratings. This was also highlighted through an exploratory qualitative analysis of the examiners' comments.

To conclude, thousands of language learners take exams such as IELTS every day, and these exams are increasingly costly. This study is important in contributing new knowledge to the area of vocabulary assessment, specifically lexical richness

measures, and their relationship to examiner ratings (furthering previous research by Read and Nation, 2002) and in working towards a model that could be used by learners worldwide to predict their IELTS score before they take the exam. This would ensure that the learners would only take the exam when ready, avoiding costly resitting in pursuit of their desired mark.

1.2 OUTLINE OF THE CHAPTERS

After this introductory chapter (Chapter 1), Chapter 2 focuses on the literature related to lexical knowledge and lexical richness. The importance of vocabulary has been highlighted by various researchers (Vermeer, 1992; Laufer and Nation, 1999; Möbarg, 1997). The size of a learner's vocabulary in a second language can be an indication of how proficient they are in that language, therefore various measures have been developed over the years in order to enable researchers to measure vocabulary knowledge. Firstly, definitions of lexical knowledge are presented with a discussion of its dimensions. Some researchers argue for the multidimensionality of vocabulary, such as Schmitt, Ching and Garras (2011), Read (2000), and Nation (2001), whereas Meara (1996) suggests that vocabulary size may be the only aspect of real importance when investigating vocabulary knowledge. The main focus here is the fact that most researchers argue that vocabulary comprises at least two dimensions: breadth and depth, and there is a difference between receptive and productive vocabulary knowledge. These are the main concepts used in this thesis. Breadth refers to the size of someone's vocabulary (how many words are known), and depth refers to a deeper knowledge of those words (how well these words are understood). Therefore, despite the many definitions used in the field (Chapelle, 1998; Henriksen, 1999; Qian, 1999) my approach to lexical knowledge is that it has at least two main dimensions: breadth and depth. This study analyses productive use of vocabulary presented in the form of essays produced by learners. The structure of Chapter 2 includes a discussion of what it means to know a word, distinguishing between knowing a word receptively and productively (Vermeer, 1992) followed by a short section on receptive knowledge tests (instruments for testing lexical diversity), such as the Vocabulary Levels Test (Nation, 1983). These were not included in this study due to the fact that they assess receptive knowledge, and only measures of productive knowledge that could be applied to the IELTS essays were used here.

After the discussion of tests of receptive knowledge the distinctions between spoken and written registers, and between academic and less academic vocabulary, are discussed.

In the next section of Chapter 2, different means of assessing lexical richness are introduced, firstly focusing on the critical presentation of different measures of lexical diversity (or variation) to explain the particular choices of measures for this study. Indices such as the well-known and very widely used TTR ratio and Guiraud's Index (Guiraud, 1960) are presented and discussed. Almost all of the measures are dependent on text length (Tweedie and Baayen, 1998), which is considered one of the main flaws of measures such as the TTR, the values of which have a negative correlation with the number of words (tokens) in essays. Therefore, the more words a user produces, the lower their TTR value because they use fewer new words the longer the text gets. Most indices of lexical diversity show a sensitivity to variations in text length (McCarthy and Jarvis, 2010), which means that the result of these measures can be affected when text length is not standardised. Then, measures of lexical sophistication (measures that go beyond counting and add a more 'qualitative' factor to the analysis, such as frequency lists) are discussed and criticised. Examples of these measures are P_Lex and Guiraud Advanced, of which a full description is given in the chapter described here. Furthermore, in this chapter, a list is provided which categorises all the tests discussed or used in this thesis into tests of either receptive or productive vocabulary knowledge.

The last, but not least important, section of Chapter 2 is a discussion of various methodological problems associated with the use of measures of lexical richness, such as its problematic definition, or the fact that different researchers operationalise the construct of lexical richness differently, hence resulting in incomparable results. Another issue discussed in this section of Chapter 2 is the method of counting words, how to deal with MWU (multiword units), the issue of small and unrealistic amounts of data, and the influence of topic and setting.

In subchapter 2.6, the construct of language proficiency (and proficiency in general) is introduced. Proficiency is a very important concept in applied linguistics, yet it is very hard to define. This issue is addressed here, along with the importance of

vocabulary in distinguishing between proficiency levels (Iwashita, Brown, McNamara and O'Hagan, 2008; Crossley, Salsbury, McNamara and Jarvis, 2011a; 2011b). The next sections in this subchapter present and discuss studies that show a link between vocabulary and various aspects of proficiency. Daller et al. (2007) suggest that the use of infrequent words could be an indicator of language proficiency. There are various other studies that support the claim that vocabulary is linked with proficiency (Pearson, Hiebert and Kamil, 2007; Laufer, 1992, 1995; Hawkey and Barker, 2004). Subchapter 2.7 consists of five sections that discuss second language testing in general, vocabulary testing, the influence of context in testing vocabulary (Schmitt in 1999 raised the question as to whether vocabulary should be tested in context or in isolation) and the last sections address the issues (associated with language testing) of reliability and validity and the notion of holistic rating (global or impressionistic rating), which is widely used but has also been criticised (Barkaoui, 2010). Even though this type of rating is easy to use, we cannot be sure what the raters actually have in mind when giving a specific score. Holistic scoring is not ideal for qualitative research in writing because it cannot identify learners' strengths or weaknesses. One of the major issues regarding the use of holistic ratings is the existence of halo effects, whereby an examiner's judgement on specific traits can be influenced by the global rating.

Subchapter 2.8 provides a discussion of the IELTS testing system and its components. This test is a well-established test of second language knowledge for learners who wish to attend British universities; it is widely used and has more than one million candidates per year. It tests four skills (reading, listening, speaking and writing). The second section of the subchapter presents findings from previous research studies related to IELTS and academic writing such as Hawkey and Barker (2004), Banerjee, Franceschina and Smith (2004) and Read and Nation (2002). The Read and Nation study is obviously central since my methodology is based on their work and this thesis develops their research. In subchapter 2.9, the relationship between vocabulary measures and teacher ratings is discussed. For instance, Crossley, Salsbury, McNamara and Jarvis (2011b) found that lexical diversity accounted for almost 50% of the variance in human ratings. Results from various other studies are presented, and variables that could act as predictors of teacher ratings, such as D or other vocabulary measures, are discussed.

Chapter 3 introduces the first small scale study of this thesis (the pilot study, Study 1). This analysis investigates two main research issues: whether measures of lexical diversity or sophistication would correlate higher with the IELTS holistic teacher ratings, and secondly to what extent teacher ratings can be predicted by a model whose variables would be various vocabulary measures. The subjects were 42 Greek-Cypriot students learning English as a foreign language and preparing for the IELTS exam. The chapter presents and analyses two sets of data, oral and written language. Guiraud (for oral) and P_Lex and tokens (for written) were the variables that accounted for most of the variability in the overall holistic scores, and these findings are further discussed in the last section of Chapter 3. In addition, the chapter discusses the limitations and issues regarding the methodology of this study, such as the low reliability of the examiners, and the potential advantage of acquiring access to a larger IELTS database, is presented.

After Study 1, which only uses measures of vocabulary breadth as predictors of teacher ratings, it is acknowledged that the research needs to be taken a step further, to look at adding other measures of vocabulary knowledge (measures of depth of vocabulary) to the IELTS model to improve its predictive validity. Depth of vocabulary is hard to define and operationalise but research (Beglar and Hunt, 1999; Qian and Schedl, 2004) suggests that the use of formulaic language (such as collocations or phrasal expressions) is an aspect of depth of knowledge. Therefore, the decision was taken to add the extra variable of 'formulaic count' to the model. Chapter 4 introduces the concept of formulaic language. Firstly, a definition of formulaic sequences (formulaic language) is provided, followed by discussions on acquisition and use, teaching and learning formulaic sequences and how to detect formulaic language in a text. Then a presentation of different examples of formulaic sequences follows. This section introduces idioms and phrasal verbs. Colocations are then introduced in a separate section (as they are the main focus of this study). Various definitions of collocation are provided, followed by a discussion on acquisition and use, and acknowledgment of the frequency factor. Furthermore, the next section presents a discussion of word lists and academic corpora, followed by a section in which the relationship of formulaic sequences and L2 proficiency (various aspects) is presented, moving on to a discussion of the relationship between formulaic language and ratings. The next section looks at various methodological problems

involved in formulaic language research. The chapter closes with a presentation of the rationale for Study 2 and an operationalisation of the study, which leads to the second, major study (Study 2) of this thesis. Formulaic sequences were operationalised by counting collocations using the PHRASE List (Martinez and Schmitt 2012) of the 505 most common phrasal expressions in English.

In Chapter 5 Study 2 is presented. The second study repeated the procedures from Study 1 with modifications, including the addition of an extra variable of the count of formulaic sequences (phrasal expressions). Therefore, the research questions were the same as Study 1 with the addition of an extra question, which asked to what extent the model could be improved by adding the extra variable of formulaic count (measure of depth of knowledge). First, an explanation is given of how this variable was operationalised by using Martinez and Schmitt's (2012) PHRASE List. This study uses a different set of data from different learners and is based on a complete re-analysis of the data collected by Turlik (2008), which is publicly available and can be found following the link:

http://www1.uwe.ac.uk/cahe/research/bristolcentreforlinguistics/researchatbcl/iclru.aspx

Turlik's methodology (his data collection procedure, and treatment of the data) is explained in the following sections with a discussion of the participants, essays, corpus, and raters' measures. After a thorough examination of existing arguments it was decided, as in the case of Study 1, that the data for Study 2 would not be lemmatized. Therefore all the derived and inflected forms of a base word would be counted as new words (types). What follows in Chapter 5 is the analysis of the data, presentation of the equipment and software used and the various calculations before moving to the next chapter presenting the data analysis.

In Chapter 6 the results of the study are presented and discussed, starting with descriptive statistics. The second section deals with regression analyses and inference. The main findings include the discovery of a strong negative correlation between the TTR and the tokens, as predicted. All measures of lexical diversity and lexical sophistication correlate with the ratings, but the number of tokens has the highest correlation of all the lexical diversity measures, and the number of types has the highest correlation of all the lexical sophistication measures. TTR, Guiraud and

P_Lex are the three measures (variables) that can explain more than 50% of the variability in lexical ratings. In addition, holistic ratings can be predicted by the same two lexical diversity measures (TTR and Guiraud), but using a different measure of lexical sophistication, Guiraud Advanced. The formulaic count did not seem to improve the model's predictive validity, but all of these findings are further discussed in this chapter.

The next section of Chapter 6 presents a discussion of Study 2. The results are presented and subsequently explained. A multiple regression leads to a new model of IELTS ratings, which is discussed, and explanations are provided of the possible reasons for the findings. After the model optioned by a multiple regression excluded formulaic count as not being significant variable, it was decided at that stage to go back to the data and re-examine them from a qualitative perspective. 30 randomly selected essays were then analysed in terms of not only how many formulaic sequences they contained, but also which ones (and with how many repetitions). These results shed some light on the previous findings because they revealed that conducting a quantitative analysis (looking at how many) was not enough. Some essays with a higher number of sequences but lower ratings seemed to have several repetitions of the same sequences, whereas essays with a lower number of formulaic sequences seemed to achieve a higher rating when these sequences were original sequences (more types than tokens). In addition, the 'qualitative' analysis showed that only 63 of the 505 phrasal expressions (formulaic sequences) in the list were used in these 30 essays, and this could indicate that some expressions may be easier than others to learn, and teachers perhaps teach them first. Quotation marks are used when referring to 'qualitative analysis' since, even though it adds a qualitative aspect to the data analysis, it stills deals with numbers. Lastly, the limitations of the study are discussed, and suggestions for further research are made. The last section of Chapter 6 offers a comparison between the present study and that of Turlik (2008), since the data (essays) and the ratings used in both studies were the same.

In Chapter 7, the holistic rating model is tested using a very small sample of real IELTS data (taken from the organisation's website). The results revealed that the statistical model was based on only a limited range of scores, therefore had only a limited predictive power and seems to underestimate the scores given for the IELTS

data. However, the examiners' comments for each essay were used for a qualitative analysis which confirmed that indeed vocabulary (positively or negatively) influences examiners' ratings as most of them mention vocabulary in the comments justifying their decisions for the given rating.

Chapter 8 provides a summary of the thesis, with concluding remarks and suggestions for further research. This chapter highlights my contribution to the field, the importance and significance of the analysis, and suggests new work that is now appropriate.

CHAPTER 2- LITERATURE REVIEW

2.1 INTRODUCTION

This chapter is a review of relevant literature regarding vocabulary knowledge, lexical richness and language testing. There is a difference that needs to be distinguished here before attempting to discuss the various definitions of lexical richness and the ways of measuring it. There is a difference between lexical richness (of a text) and the many ways measuring it and lexical knowledge (and the many dimensions of it). There are many definitions used in the field of second language acquisition and vocabulary research. The topics discussed in this chapter are the following: vocabulary acquisition and use, definition of vocabulary-lexical knowledge, dimensions of vocabulary-lexical knowledge (breadth and depth) and the difference between receptive and productive vocabulary. This discussion also highlights the relationship between productive vocabulary knowledge and the 'evidence' of this as vocabulary produced in a text. Then follows a discussion on the distinction between spoken and written registers and academic (and less academic) writing. Furthermore, lexical richness definition and problems, measures of lexical richness (lexical diversity and lexical sophistication measures) and their advantages and disadvantages and methodological problems when attempting to measure lexical richness are presented. In addition, a discussion on the construct of language proficiency is provided with a presentation of studies regarding the relationship between vocabulary and various aspects of proficiency. A short discussion on language testing and scoring follows, and the IELTS exam components are presented and explained. Lastly, studies that show the link between vocabulary measures and teacher ratings are discussed.

2.2. VOCABULARY ACQUISITION AND USE

According to Schmitt and Zimmerman (2002), vocabulary acquisition and use has an integrated/incremental nature. This means that vocabulary acquisition is a gradual procedure and language teaching and learning programmes should include recycling/repetition of vocabulary in their curriculum as the learners need to be exposed as much as possible to the targeted vocabulary. Hatch and Brown (1995) also support the idea that vocabulary learning is not a straightforward procedure that

can be achieved just by memorising a list of words (word lists) but a more complicated one which is accomplished through constant vocabulary use. When referring to the organisation of language in the mind, it is suggested (Carter, 1987) that, besides the conceptual memory there exists a mental lexicon, which stores a plethora of information about different words (phonological, morphological, syntactic, pragmatic information etc.) The acquisition of words, in both L1 and L2, is not a procedure involving acquiring words as single entities, but as L2 'labels' of concepts, which form larger domains of knowledge and form the network of our knowledge of the world. The way a word is pronounced or heard is the L2 label of a word, while the concept is everything else that is linked to the word (meanings, associations, ideas and images). These previous studies support the idea of a quite complicated procedure regarding the way vocabulary is acquired and stored in our brains. This definitely has implications on testing/measuring vocabulary research as presented in following chapters.

2.3 VOCABULARY/LEXICAL KNOWLEDGE

As Vermeer states 'knowing words is the key to understanding and being understood' (Vermeer, 1992: 147). Vocabulary knowledge is also an important aspect of language assessment and is regarded to be one of the main aspects of language competence (Grabe, 1991; Frederiksen, 1982). Laufer and Nation (1999:38) state that 'learners at a higher level of language knowledge know more words', and according to Möbarg, 'vocabulary is arguably the most important aspect of language learning' (Möbarg, 1997:201). According to Turlik (2008) real vocabulary suggests a continuum (consists of different learning stages). In addition, when referring to vocabulary knowledge a distinction must be made between receptive and productive knowledge (see following Section 2.3.4 for tests of receptive and productive knowledge-discussion).

2.3.1 Dimensions of lexical knowledge

Many researchers (Schmitt, Ching and Garras, 2011) have in recent years realised the need to look at vocabulary's multidimensionality (size and depth). Other researchers that argue for the multidimensionality of vocabulary knowledge are Read, Wesche

and Paribakht, and Qian. 'Read (1989), Wesche and Paribakht (1996), and Qian (1999) state that vocabulary knowledge should at least comprise two dimensions, vocabulary breadth, or size, and depth, or quality, of vocabulary knowledge' (Qian and Schedl, 2004:28). Moreover, Chapelle (1998) proposes a four dimensional definition of vocabulary that consists of vocabulary size, knowledge of word characteristics, lexicon organization, and processes of lexical access. Furthermore, Henriksen (1999) suggests a 3-dimensional model with the main components being 1) precision of knowledge, 2) depth of knowledge, and 3) receptive and productive knowledge (see more on receptive vs. productive knowledge in Section 2.3.4). Daller et al. (2007), metaphorically speaking, also argue for a three dimensional 'lexical space', which consists of breadth, depth and fluency but as Turlik (2008) suggests this definition might be difficult to operationalise because it is quite problematic to define criteria such as 'well known', 'depth', 'breadth' and 'fluent'.

An empirical study by Qian (1999) which investigated the relationship between the dimensions of breadth and depth of vocabulary and reading comprehension in ESL found that two tests, the Vocabulary Levels Test (VLT) (Nation, 1983;1990) which is a test of breadth/size of vocabulary and the Words Associates Test (WAT) (Read, 1993), which is a test of depth, correlated significantly and closely and states that these two aspects of vocabulary knowledge are equally important, as they overlap one another and are interconnected (see Section 2.4.4. and 2.4.5 for descriptions of these two tests). This finding was also supported by Akbarian (2010) who investigated the relationship between breadth and depth of vocabulary for Iranian learners of English using the same two tests (VLT and WAT). There was a strong correlation between the two tests and 'the findings suggest that vocabulary size and depth might be accounted for by the same factors, especially as the learners' proficiency increases' (Akbarian, 2010:391). Other researchers that suggest that there are two dimensions of vocabulary, size and depth but they seem to be highly correlated, are Bogaards and Laufer (2004), and Milton (2009). Milton (2009) argues that there may be no distinction between those two dimensions as they are very closely related. Vermeer (2000; 2001) supports this idea (especially in lower levels) and Verhallen and Schoonen (1998) also support the fact that breadth seems to be correlated with depth, an idea also proposed by Read (2004) who suggests that there is evidence that these two dimensions are not opposites but are closely related. Even

though there is much evidence of the high correlation between the two vocabulary knowledge dimensions, some researchers do seem to deal with these as contrasting aspects of vocabulary knowledge. As a researcher I believe that even though studies suggest that these two dimensions are highly related there are also some aspects that are representative of one and not the other. Besides Akbarian (2010) admitted that the WAT may not even be a depth test at all (but a breadth one in disguise) - an idea supported by Milton (2009) - and that would explain the high correlation between the two dimensions. Therefore, this is one of the main reasons that after the pilot study of my own research a second study was conducted in which an aspect of depth of knowledge (collocations) was investigated to see if it could help with the construction of my predictive model (this will be discussed in detail at the end of Chapter 4). Laufer, Elder, Hill and Congdon (2004) propose that both size and strength of vocabulary are equally important in vocabulary testing: 'In sum, it appears that for diagnostic purpose we need separate estimates of both size and strength to fully understand the degree of a learner's vocabulary knowledge' (2004:224). Therefore, knowing a word is multidimensional and has many degrees of knowledge such as receptive and productive knowledge or looking at specific aspects of vocabulary such as collocations (see Section 2.3.2 for a discussion on this issue). To get a better idea of someone's vocabulary we need to take various measures into account (Laufer and Nation, 1999.) Brown's (2011:83) study suggests that 'a more rounded view of vocabulary knowledge needs to be adopted by material writers, and argues for an approach in which items are revisited regularly as different aspects of vocabulary knowledge are introduced'. Nation (2001) suggests that there are at least nine aspects of vocabulary knowledge.

Singleton (1999) criticises the approach adopted by researchers such as Laufer and Nation (1999), who treat and count vocabulary as a single phenomenon independent from grammar, text or discourse. Singleton suggests that investigation of factors besides size and growth of vocabulary need to be added. (Read and Chapelle, 2001) This suggestion contrasts Meara's (1996:45) statement, which argues that 'vocabulary size is probably the only dimension of any real importance as long as we are dealing with a small lexicon'. I would agree with most of the researchers and suggest that when investigating vocabulary more than one dimension/aspect should be taken into account. The most common view and one of the main ideas in my study

is that vocabulary knowledge has at least two main dimensions: breadth and depth. Milton (2009) suggests that it is very hard to measure depth because the construct is very unclear and ambiguous, but proposes that one way to address depth is by measuring the aspects of depth, such as idioms or collocations, separately (see Study 2). This idea is also proposed by Schmitt (2010:13) who describes that depth of knowledge can be conceptualised by overall proficiency or by breaking it to components (such as spoken form, collocations, meaning etc.) which he calls the 'component' or 'dimensions' approach.

To summarise, when researchers attempt to analyse vocabulary they should do so by approaching their studies from two perspectives: receptive vs productive vocabulary and size vs depth. For my study, I approach the operationalisation of depth from a 'components' perspective (as suggested by Schmitt, 2010) which means that instead of using the general definition of how well a particular word is known I take into account various components, one of which is collocational use (see Chapter 4 for further discussion). Therefore in my case I operationalise the term by the use of collocations (a person knows a word well if they know the word's collocates). Since according to Akbarian (2010:393) the 'construct validity of depth is therefore challenged' one of the common approaches researchers use, including myself, is to test some of the aspects (such as collocations), that constitute vocabulary depth, separately and assume this aspect will represent ability in the whole spectrum (of vocabulary depth). Breadth is operationalised by measuring learners' vocabulary size (using various diversity and sophistication measures). I am measuring learners' productive vocabulary since essays (produced by the learners) are used for the analysis in the present study.

In the last twenty years vocabulary research has grown due to technological advances and the availability of large corpora. '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, 2001:180). In recent years investigation into lexical richness has been carried out by several researchers (Vermeer, 1992, Laufer and Nation, 1995, Malvern and Richards 1997, 2002 etc.).

There are many approaches to calculating or measuring lexical richness, for example measures of lexical diversity (indices and tests) and measures of lexical sophistication (Malvern et al., 2004) which will be discussed in detail in following sections.

2.3.2 What does it mean to know a word?

One of the main discussions regarding vocabulary knowledge is what constitutes a word. There are many different definitions in the field (Cronbach, 1942; Richards, 1976; Nation, 1990; Carter, 2000) and this is one of the challenges met when researchers wish to conduct research regarding vocabulary knowledge/lexical richness measurement. A small discussion is found here but this will be discussed further in Section 2.5 (Methodological problems when measuring vocabulary).

Richards (1976) argued about seven aspects of word knowledge which include syntactic behaviour, associations, semantic values, different meanings, underlying form and derivations. Nation (2001:23) states:

> 'Words are not isolated units of language, but fit into many interlocking systems and levels. Because of this, there are many things to know about any particular word and there are many degrees of knowing.'

Nation provides an analytical table regarding what is involved in knowing a word (Nation, 2001:27). This includes both receptive and productive knowledge of the following: Form (spoken, written and word parts), Meaning (form & meaning, concept & referents and associations) and Use (grammatical functions, collocations and constraints on use). According to Nation (1990) collocations and frequency are both dimensions of what constitutes a word. Please see Table 2.1 below:

Table 2.1: What is involved in knowing a word? (Nation, 2001: 27)

Form	Spoken	Receptive	What does the word sound like?
		Productive	How is the word pronounced?
	Written	Receptive	What does the word look like?
		Productive	How is the word written and spelled?
	Word parts	Receptive	What parts are recognizable in this word?
		Productive	What word parts are needed to express this meaning?
Meaning	Form & meaning	Receptive	What meaning does this word form signal?
		Productive	What word form can be used to express this meaning?
	Concept & referents	Receptive	What is included in the concept?
		Productive	What items can the concept refer to?
	Associations	Receptive	What other words does this make us think of?
		Productive	What other words can we use instead of this one?
Use	Grammatical functions	Receptive	In what patterns does this word occur?
		Productive	In what patterns must we use this word?
	Collocations	Receptive	What words or types of words occur with this one?
		Productive	What words or types of words must we use with this one?
	Constraints on use (register, frequency…)	Receptive	Where, when and how often would we expect to meet this word?
		Productive	Where, when and how often can we use this word?

According to Vermeer (1992), knowing a word involves knowing the concept behind that word. Vermeer states that there are two 'ways' of knowing a word. Words can be known receptively (known in a context only), and productively. Learners' receptive control of new words precedes their productive control and the size of the receptive vocabulary is larger than the size of the productive vocabulary. This is also

supported from studies by Fan (2000), Laufer (1998), Waring (1997) and Webb (2008). Vermeer argues that, due to the fact that it is hard to define what a word is or what 'to know a word' means, it is very hard to indicate the size and growth of vocabulary in children and compare results from different studies. When conducting research to measure people's vocabulary it is hard to distinguish between receptive and productive knowledge (it is very hard to know if the person really 'knows' all the words). Vermeer states that it is very hard to operationalise absolute size and growth of vocabulary due to the fact that many of the measures available are neither valid nor reliable. Laufer (1997) states that one of the factors which determine whether someone knows a word is by knowing its common collocations (see more on collocations in Chapter 4). Even though it is not sufficient to view vocabulary as single words (as lexical knowledge is much more complicated and consists of many dimensions as will be discussed below) most researchers, for practical and testing reasons, use the definition that anything between two spaces is a word.

2.3.4 Receptive vs. Productive Vocabulary and List of Tests

When addressing vocabulary knowledge it needs to be acknowledged that there is a distinction between receptive and productive vocabulary. Receptive vocabulary refers to the amount of words a learner can handle in reading or listening situations, whereas productive vocabulary knowledge refers to all the words that are available when a learner is required to speak or write in an L2 (Daller et al., 2007). Meara and Fitzpatrick (2000) also state that productive vocabulary is written or spoken vocabulary produced by the learner. The terms active and passive are also sometimes used to refer to productive and receptive skills (Meara, 1990). Read (2000) presents a very detailed analytical table that was provided by Nation (1990) - see table 2.1 above- as to what is considered receptive and productive vocabulary knowledge and in very simple terms he explains that 'it is the difference that we are all familiar with between being able to recognise a word when you hear or see it and being able to use it in your own speech or writing' (Read, 2000:26). Therefore, following Read's definition, any vocabulary produced by a learner in a text (composition) or sample of speech will be treated and will be considered to be under the term 'productive vocabulary knowledge'. Besides, as Nation (2001:25) suggested: 'Productive vocabulary use involves wanting to express a meaning through speaking or writing

and retrieving and producing the appropriate spoken or written word form'. According to Schmitt (1998a), instead of differentiating between knowing a word receptively or productively it is more appropriate to say that words (or aspects of word knowledge) are known to different receptive and productive skills (Schmitt 1998a cited in Turlik 2008: 37). Researchers such as Melka (1997) support the theory that the relationship between receptive and productive mastery of vocabulary is a continuum where the first precedes the latter (we first learn words receptively and then productively). This is also supported by Schmitt (1994) who states that the distinction between receptive and productive vocabulary (and vocabulary testing) is more of a continuum rather than a dichotomy. According to Meara and Fitzpatrick (2000), receptive vocabulary knowledge is larger than productive vocabulary, and it is more difficult to measure and rate productive vocabulary than receptive vocabulary knowledge. Webb (2005) also claims that when it comes to learning vocabulary there is also a difference between receptive and productive learning but not much research was conducted to show the differences between the two.

There are various tests of receptive and productive vocabulary knowledge. Below is a list of measures/tests discussed in the thesis, categorised either as measures of receptive or productive vocabulary knowledge.

List of Measures

Receptive	Productive
VLT (Vocabulary Levels Test)	NDW (Number of Different Words)
Revised form of VLT	TTR (Type-Token ratio)
VKS (Vocabulary Knowledge Scale)	D
WAF (Word Associates Format)	Guiraud Advanced
Yes/No Test	Lex30
	LFP (Lexical Frequency Profile)
	P_Lex

All the receptive tests listed and one of the productive list (Lex 30) are all single item tests of learner knowledge whereas the rest of the productive tests are all measures of texts. According to Nation (2001), one can test productive written vocabulary knowledge by either using a discrete-point vocabulary test or by analysing the vocabulary of learners' written compositions (which is the analysis used in the present study).

2.3.5 Distinction between spoken and written registers

Since my pilot study (Study 1) involved the collection and analysis of both oral (spoken) and written data, it is appropriate to highlight the differences in the nature of these two registers. Therefore, a small discussion is included here in order to present some of these differences. According to Nation (2001), speaking requires a smaller vocabulary size than writing probably due to formality and topic differences (between the two registers). There are of course some characteristics that are representative of one register and not the other. For example, some vocabulary items such as hedges, greetings and softeners are more likely to occur in spoken than written language (Nation, 2001). Schmitt (2010) provides an example of some words more common in speech such as *yeah* and *okay* and an example of words more commonly found in writing such as *thus* and *political*. In addition, according to Schmitt (2010:14), 'the frequencies of lexical items differ considerably between spoken and written discourse.' There are major differences between using written or spoken corpora (Shin, 2007). The collocations used in oral and written speech are considerably different, but it has to be said that collocations are found and used more often in oral speech than written speech (see more on collocations in Chapter 4).

Another aspect that distinguishes between the two registers is lexical density .This is one of the dimensions used in Read's definition to describe lexical richness. It is a dimension that discriminates between written and spoken registers (Malvern et al. 2004) and is more appropriate (according to Read) for spoken language. As Fatahipour (2013:63) suggests 'relation between lexis and writing is not straightforward and depends on other factors, among which assessment issues and task choice play a crucial part'.

Regarding measuring lexical richness, analysing written texts is more common than using spoken language. However, Daller van Hout and Treffers Daller (2003), Malvern et al. (2004) and Treffers-Daller, Daller, Malvern, Richards, Meara and Milton (2008) are some of the researchers that measure lexical richness via speaking while most researchers use written texts. G Yu (2009- See Section 2.4.3) is one of the researchers that measures lexical diversity in both speaking and writing of the same students.

To summarise, the differences between oral and written registers highlighted in this section could provide some explanation later in the thesis regarding different results for the two datasets in Study 1.

2.3.6 Academic Writing

This thesis investigates the relation between measures of lexical richness and IELTS ratings through the analysis of academic text. Academic writing should be treated differently than writing for other purposes. As already discussed in the previous section, there are certain characteristics that are representative of either written or oral registers. If we focus on just written data, there is also a distinction between different genres. For example, academic writing differs and consists of different lexical characteristics than less academic writing. Academic vocabulary consists of both high frequency words and technical vocabulary but also non-high-frequent words which are common across academic disciplines (Schmitt 2010:78). The AWL is the best available list of academic vocabulary (see more on AWL in Chapter 4). The nature of academic writing influenced the choice of specific measures (especially lexical sophistication measures that are based on frequency lists) and could be an influential point for some of the results of Study 1.

2.4 LEXICAL RICHNESS

The size of someone's vocabulary (the number of words that a person knows) can be an indication of how proficient they are in that language. Therefore, researchers in second language acquisition and assessment have argued that it might be necessary to find a way to measure lexical richness (in a text as indication of the vocabulary

knowledge of the originator of a text) in order to understand the level of learners of that language (Read, 1993; Laufer et al., 2004) (see below for definition).

2.4.1 Definition: a single or multi -dimensional model?

Extensive work has been carried out on forming methods of measuring lexical richness. Before addressing the various methods developed to measure lexical richness, it is important to provide a definition of this term. According to Malvern et al. (2004), for some researchers the terms lexical richness and lexical diversity are synonymous, but in this study I adopt a different approach. Of course there is more than one definition according to different researchers and there is no established definition of lexical richness, but I will start with the definition given by Read, of which I will be testing two of its dimensions in my research (lexical variation, also called lexical diversity (see below) and lexical sophistication).

According to Read (2000:200-5) lexical richness has the following dimensions:
1) 'Lexical variation' which refers to the number of different words used in a text (not the total number of words)
2) 'Lexical density', which is the ratio between content words and function words
3) 'Lexical sophistication', which is the use of 'rare' or infrequent words
4) 'Number of errors', which means that someone with a high vocabulary level will only make a few minor vocabulary errors.

These errors include choosing an inappropriate word to express an intended meaning, words that do not have the correct form or the correct style and words that would be grammatically incorrect when positioned in certain places in sentences. There is a similar approach by Laufer and Nation (1995) discussed in Section 2.4.5.

For my study I use two of Read's components as variables (lexical variation/ diversity and lexical sophistication) and use lexical richness as a cover term which includes both aspects. This definition of lexical richness in my study is adopted by Malvern et al, 2004 which suggest that when referring to the term lexical richness one actually refers to lexical diversity (or variation) and lexical sophistication.

The term 'lexical variation' is interchangeable with the term 'lexical diversity', the latter term will be used throughout this study due to the fact that it is more commonly used (Malvern et al., 2004).

2.4.2 Lexical diversity measures

Measuring vocabulary (or counting words) dates back a long time (Thompson and Thompson, 1915; Fries and Traver, 1960; DeRocher, 1973; Nation and Waring, 1997). Lexical diversity is very important for testing in various fields such as neuropathology, stylistics, and language acquisition (McCarthy and Jarvis, 2007). A very simplified definition of lexical diversity would be the range and variety of vocabulary a learner uses in their speech or writing (McCarthy and Jarvis, 2007). However, defining lexical diversity is very challenging and quite problematic as will be discussed in a following section (see Section 2.5 titled Methodological problems when measuring vocabulary). Many indices were developed in order to try and measure learners' lexical diversity, yet a fully valid and reliable lexical diversity measure has proven to be elusive (Jarvis, 2002). It is very important for a reliable and valid measure to be found in order for researchers to be confident in the conclusions they draw.

2.4.3 Indices/Measures Based on Mathematical Models or Ratios

The basic measurement of lexical diversity is simply counting the number of tokens and types. The word 'tokens' refers to the sum of all words in the text (total number of words), where the word 'types' refers to each individual word (different words). One of the simplest methods of measuring someone's vocabulary size is the NDW, or number of different words from a sample, which is used to calculate the range of a learner's vocabulary. This measure gives only a single value and, according to Malvern et al. (2004), has some disadvantages as it is strongly related to sample size. This means that larger texts will acquire higher values. Another method used to calculate lexical diversity is the use of ratios. One of the most common measures of lexical variation/diversity is the type-token ratio or TTR first introduced by Johnson (1944). As was previously mentioned, tokens represent all the words in a text, and types are the different words in a text. So, as Malvern et al. note, 'When a word is repeated, then there will be two tokens (or more) of one type' (Malvern et al.,

2004:19). To calculate the TTR it is necessary to divide the numbers of types by the number of tokens:

$$TTR = \frac{\text{Types}}{\text{Tokens}}$$

A value between 0 and 1 is given when calculating TTR, and the higher the value, the greater the lexical richness of the text. Here it needs to be clarified that it is the lexical richness of the text that the TTR gives, not of the author. Authors with a large vocabulary are able to write simple texts (e.g. for stylistic reasons). The low lexical richness of the text is then not automatically an indication of poor vocabulary knowledge of the author. These two things are not the same. According to Malvern et al. (2004), calculating ratios is a better measurement than simple raw values. Malvern et al. (2004) state that research shows that this measure is flawed as it can be affected by the size of the sample in a similar way to the NDW. Higher values can then be acquired from shorter texts, and with larger texts the TTR will give you lower values. Even for a text written by the same author the values for the TTR decrease with increasing text length, as previously mentioned. If the texts are written by different authors, then larger texts can be an indication of a higher proficiency and there is nothing wrong with them getting higher values for an index. The point is that larger texts get systematically higher values even if they are written by the same author. A text from the same author gets a lower value at the beginning and a higher value at the end although the proficiency/vocabulary knowledge of the author does not change. Even though there have been various attempts of standardisation,- attempts to standardise the size of samples, number of tokens (Klee, 1992; Thordardottir and Weismer, 2001) - there are still problems with this measure. Therefore, even though this is the most obvious and simplest way of measuring lexical diversity, it is flawed (Duran, Malvern, Richards and Chipere, 2004).

Malvern and Richards (1997) explain that, even though TTR is one of the most common measures of lexical diversity, there are many problems with it, the biggest problem being the fact that TTR is not constant and decreases in parallel with increasing text length/number of tokens. This means that the more a speaker talks (or writes), the greater the possibility of repetitions (they run out of new words with increasing text length because the vocabulary size of every speaker/writer is finite).

Therefore, the main weakness of TTR is its sensitivity to text length (Read and Nation, 2002). This problem could be explained by Heaps' law (1978) - see Glossary of Terms and Abbreviations- which suggests that with the increase of tokens in a text, the number of types falls. Therefore, the more tokens, the less types will be produced in a learner's speech or text (because types will re–appear as tokens if the learner repeats them).

To make these measures valid, a standardisation is needed. So, to compare different TTR ratios, it is necessary to use the same number of tokens for each person, in order to make the ratios comparable (Malvern et al., 2004:24-5). Van Hout and Vermeer (2007) support the above statement by stating that 'plain Type-Token Ratios can produce erratic outcomes, especially when the numbers of tokens vary substantially between the texts to be compared' (Van Hout and Vermeer, 2007:93). The researchers comment on the fact that it is remarkable that even though the TTR is proven to be erratic it is still widely used in several studies (including my study- see below for justification). Despite these methodological problems I included TTR in my research design because it is a widely used measure (see Methodology section).

Another traditional and one of the first developed measures is the MSTTR. MSTTR (Mean Segmental Type-Token Ratio) is a measure which was first recommended by Johnson (1944). Malvern et al. (2004:25) describe that the calculation of MSTTR can be done by:

> 'Choosing a given standard number of tokens, sufficiently small for a number of different sub- samples of that size to be taken from the smallest language sample in the data set. Each transcript is then divided up into segments of the given length and the TTR calculated for each. MSTTR is the average over all sub-samples.'

MSTTR is not a function of sample size because the size of the segments (whose TTRs are calculated) are averaged therefore ensures higher reliability. Even though it is obvious that MSTTR is an improved version of NDW and raw TTR, it is still not considered the best measure of lexical diversity due to various problems associated with it (Malvern et al., 2004). Some of the main problems according to Malvern et al. (2004) associated with this measure are the following: non comparability of MSTTRs calculated by different sizes of standard segment, not suitable for very short texts

because they give distorted results and loss of data in cases that transcripts cannot be divided exactly into standard-sized segments. The researchers state that their measure D (see below) overcomes all these problems, and this justifies my decision for its inclusion (instead of the MSTTR) in the present study.

For many decades researchers have tried to improve the existing indices in an attempt to overcome the text-length dependency weakness. There have been attempts by various researchers to transform TTR, including Guiraud (1960) and Carroll (1964), who proposed the 'Corrected TTR'. Guiraud proposed the 'Root TTR' to solve the problem of TTR and sample size. Mainly, what the researchers were trying to do was to try to create a constant and overcome the fact that TTR falls with increasing text (Malvern et al., 2004). Guiraud assumed that 'the fall is proportional to the square root of the token count' and the measure he proposed was TTR multiplied by the square root of N (Malvern et al., 2004:26).

$$RTTR = \frac{V}{\sqrt{N}} = \frac{\sqrt{N}}{N}V = \sqrt{N}\frac{V}{N} = \sqrt{N} \times TTR$$

Logarithmic transformations of the TTR (Herdan, 1960) were also proposed to overcome its flaws which are mentioned above. Herdan's Index or LogTTR is calculated by dividing the logarithm of tokens by the logarithm of types:

$$Log\ TTR = \frac{\log V}{\log N}$$

Maas (1972) proposed another index that uses a logarithmic transformation of types and tokens. McCarthy and Jarvis (2010) reported that Maas (1972) was the index from the log correction approach that was the least affected by text length. Maas index is calculated using the following formula:

$$MAAS = \frac{\log N - \log V(N)}{\log^2 (N)}$$

According to Malvern et al. (2004), all these measures were tested over a number of years which brought researchers to the conclusion that most of them are still quite problematic because none of them seem to overcome the sample size problem. All the mathematical transformations of TTR were found by Vermeer (2000) to be unsatisfactory in terms of their validity or reliability. However, Vermeer (2000) claims that Guiraud is better than TTR. As already mentioned above, according to G Yu (2009), even though there were various attempts to develop measures of lexical diversity, the TTR is still the most widely used (despite its flaws).

To solve the problem of text length dependence, Malvern and Richards (McKee et al, 2000) developed a mathematical model of lexical diversity and introduced D, which is the single parameter of a function that models the falling TTR curve. The fall of the TTR curve is less steep for essays with a greater lexical richness than for essays with lower lexical richness and therefore, 'the value of D determines the height of the curve and therefore measures the diversity of vocabulary'. (Malvern et al., 2004:189). According to G Yu, 'the higher the D, the greater the diversity of a text' (2009:239). The minimum sample size requirement to compute a valid D is 50 words. However, as Van Hout and Vermeer (2007) state, all kinds of measures (even D, which was proposed to offer a solution to previous problematic measures) seem to have reliability or validity issues and suggest that the TTR can sometimes be a better measure than D in terms of concurrent validity. On the contrary, Malvern and Richards's (2002) research, established that D is a valid measure of vocabulary diversity. It can be computed using the *vocd* command with the CLAN software (MacWhinney, 2000) to analyse language data and measure someone's vocabulary. This command uses random or sequential sampling to calculate lexical diversity (McCarthy and Jarvis, 2007). Also, D as a measure seems to be text length dependent and this will be discussed in the following paragraph.

Hoare (2000) wanted to examine whether D is a better measure of vocabulary richness than TTR, and sought to uncover whether D is dependent on text length. In his study, calculations of the TTR and D from EFL students' oral stories were compared. Two different groups of non-native speakers who were learning English were used, one intermediate and one advanced level group. The subjects were asked to describe two pictures (the first was just a picture and the second was a picture

story). Then the transcripts were transcribed into CHAT (MacWhinney, 2000) format, and using CLAN tools the TTR and D for each subject was calculated. The results showed that the mean D figure for the advanced group was significantly higher than the figure for the intermediate group which was an expected result. Even though the TTR produced a similar result, it had to be discounted because of the effect that the shorter utterance length had on the TTR score. Therefore, it was suggested that D is a more accurate way of measuring lexical diversity than TTR. However, the important outcome of Hoare's study is the fact that he showed that D is text length dependent (but not for texts written by the same author). This needed to be examined further and could not be generalised due to the fact that this was a small scale study.

According to G Yu (2007), lexical diversity is used as a part of the rating scales of many widely used tests such as IELTS, TOEFL and MELAB (Michigan English Language Assessment Battery). Specifically in IELTS, all writing and speaking samples are rated for their 'lexical resource'. According to G Yu, lexical diversity is also used for automated writing and speaking scoring. 'It seems that lexical diversity had been widely assumed as an important quality indicator of test performance' (G Yu, 2009:237). D is a good measure of academic performance (G Yu, 2009) and has many methodological advantages (Jarvis, 2002; Malvern et al, 2004). Crossley et al. (2011a) state that D is indeed a good predictor of academic performance (this will be further discussed in Section 2.9). According to G Yu (2009), the use of similar measures such as D and TTR may provide contrasting findings. However, this does not necessarily mean that the one is better than the other, but the author does urge us to look at lexical diversity from various perspectives. G Yu (2009) also states that lexical diversity can be affected by various non-linguistic factors such as stress, anxiety (Howeler, 1972) or anticipation of being evaluated (Jarvis, 2002).

However, McCarthy and Jarvis (2007) discussed and criticised the validity of *vocd* (which produces D) as a measure of lexical diversity. They argue that a fully valid and reliable measure of lexical diversity has not yet been found. They criticise the use and importance of *vocd* as the standard/norm for measurement of vocabulary in many areas such as stylistics, neuropathology, language acquisition and forensics because often researchers get misleading, questionable results. One of their main arguments is whether it really measures what is supposed to measure, they basically

criticise the wide use of *vocd* as a tool of lexical diversity measurement. They state that even though D (*vocd)* is a better measure than others, it should still be used with caution because it is also sensitive to text length (this is also supported by Hoare-previous page). The authors claim that Malvern and Richards, the creators of *vocd,* stated that the upper limit of the index is a non-specific 'few hundred words' without clearly defining this. Why is this relevant? Because as McCarthy and Jarvis rightly argue, researchers will always want to investigate texts of more than a few hundred words (longer texts) and will always want to compare essays of different sizes.

Moreover, McCarthy and Jarvis (2007) state that researchers should test measures of lexical diversity (LD) against each other. Therefore, they tested *vocd* against 13 other measures of lexical diversity. The thirteen measures that were rivals of D in this study were: RTTR (Guiraud, 1960) and CTTR (Carroll, 1964), which are square root correcting measures. U (Dugast, 1978), SS (Somers, 1966), Maas (Maas 1972), and RK (Rubet's K, Dugast, 1979) which are log correcting measures. M (Michea, 1969), S (Sichel, 1975), and K (Yule, 1944) as measures that regulate frequency of occurrence of types, and W (Holmes and Singh, 1996; Bucks, Singh, Cuerden and Wilcock, 2000). The original calculation of D was then added, calculated using 2 ways (Malvern and Richards, 1997, and Jarvis, 2002), and the last measure was the traditional Raw TTR. After a Pearson correlation it was found that all 14 measures correlated significantly with the text length, suggesting that even though *vocd* is a good measure of LD, it still had not overcome the text length dependency problem older indices had.

Overall, McCarthy and Jarvis (2007:461) are very critical of *vocd'*s reliability and construct validity. The authors do not support the use of *vocd* as an 'industry standard for measuring Lexical Diversity' due to the text length sensitivity. They support that a definition of the construct of lexical diversity is also required, and question whether just one measure (one single index) is enough to encompass the construct of lexical diversity.

Fatahipour (2012) investigated the validity of various lexical richness measures. He investigated the validity of D and Guiraud which are lexical diversity measures (and Guiraud Advanced but this is not relevant in this section). Fatahipour (2012) is quite

critical regarding the use of D as a valid measure of lexical diversity as it did not produce (in his study) high significant results (there was a correlation of D and general language ability- measured by the VLT- but not a highly significant one). His results showed that Guiraud proved to be a better measure of lexical richness since there was a highly significant correlation between Guiraud and language ability (VLT scores).

There are many different measures that can be used for measuring learners' lexical diversity, all with their advantages and disadvantages. Overall, most researchers (Silverman and Bernstein Ratner, 2000; Owen and Leonard, 2002; Malvern et al., 2004) seem to agree that D (even though it has its drawbacks) is a good predictor of ratings and can be used as an indicator of academic performance. However, as Malvern et al. (2004) suggest, a combination of various measures and aspects should be taken into account because 'a single, perfect measure of lexical diversity fit for all research purposes may be just like the Holy Grail' (Malvern et al, 2004:3).

2.4.4 Test/instruments (receptive knowledge tests)

As already mentioned in Section 2.3.4, vocabulary can be analysed in terms of receptive and productive vocabulary. This small section here is about various tests of receptive vocabulary knowledge. Some of the tests are listed here to give an overview of what is available for measurement of what is thought to be receptive vocabulary knowledge but I will not expand on this issue as my focus is different (I focus on productive vocabulary knowledge- essays written by learners).

Apart from the various indices for measuring productive vocabulary, there are also tests that have been developed to measure receptive vocabulary knowledge. The Vocabulary Levels Test has been the centre of attention for many years (Read, 2000) due to its use as a placement tool and a measure of learners' vocabulary and size. The Vocabulary Levels Test by Nation (1990) is not a measure of vocabulary sophistication (depth of knowledge), but rather a measure of the learners' knowledge of common word meanings at various levels (2000, 3000, 5000, 10000 and University Word Levels). It comprises of these five levels based on word frequency (Beglar and Hunt, 1999). The Vocabulary Levels Test (Nation 1983, 1990) is not, according to

Nation (2001), a test to be used as a measure of vocabulary size, it is a measure of written receptive vocabulary size, so it would not provide much information on how the test words could be used in speaking or writing tasks (Nation 2001 cited in Beglar, 2010). However, it is a diagnostic test in terms of vocabulary teaching since it was created to assist vocabulary teaching in learning programmes. This idea seems to be reinforced by Cameron (2002) who states the usefulness of the Levels Test as a research and pedagogic tool (for receptive vocabulary size). The Vocabulary Levels Test was criticised by Read (1993) because it presents words in isolation (see Section 2.7.3 regarding the influence of context in testing vocabulary). Beglar and Hunt (1999) revise and validate through their study the 2000 Word Level and University Word Level Vocabulary tests which are both components of the Vocabulary Levels Test developed by Nation. The authors propose two new forms for each test which make the tests more reliable. Another test of receptive vocabulary knowledge is the Yes-No Test (which was developed by Meara and Buxton, 1987). Receptive vocabulary knowledge is measured by asking the participants whether they know a word or not (pseudo-words are used in this test format to control for guessing). Mochida and Harrington (2006) investigated the Yes-No Test, and suggest that it is a valid measure of receptive vocabulary knowledge. According to Mochida and Harrington its checklist format makes it quick and easy to use.

As I have already mentioned at the start of this sub-section, none of these instruments were used for measuring vocabulary knowledge in this study due to the fact that these tests are measures of receptive knowledge. In my study all the measures used were of productive vocabulary knowledge that could be applied to measure the vocabulary in the IELTS writing transcriptions.

2.4.5 Lexical sophistication measures

Vermeer (2000) suggested that it is not enough to deal with numbers alone, and that adding frequency of words (difficulty of words) in a model of lexical diversity would make it more valid (this is however quite contradicting because frequency data are also numbers). It was suggested by Van Hout and Vermeer (2007) that existing lexical richness measures could be improved by adding a frequency factor. Martinez and Schmitt (2012) also claim that the issue of frequency is at the forefront of

research. Schmitt (2010:13) states that: 'Frequency is one of the most important characteristics of vocabulary, affecting most or all aspects of lexical processing and acquisition.'

All measures/indices from previous section 2.4.3 are quantitative measures, as all of them are based on the relationship between types and tokens. However, there are also measures of lexical sophistication that are based on the use of frequency lists (i.e. they look at the advanced words) such as the lexical profile 'LFP' (Laufer and Nation, 1995), P_Lex (Meara and Bell, 2001) and Guiraud Advanced (Daller, Van Hout and Treffers-Daller, 2003). These measures focus on different aspects of lexical richness because they make a distinction between infrequent and basic words, so they are used to measure lexical sophistication (Daller and Phelan, 2007). Hellman states that (2011:178) '...the primary source of adult vocabulary growth is exposure to low-frequency vocabulary in a wide range of texts…' which supports the focus on frequency when measuring lexical richness.

Word frequency can also be particularly useful in terms of vocabulary teaching and learning. According to Daller et al. (2007), frequency and vocabulary learning are closely associated. Laufer and Nation (1999) state that the distinction made between high frequency and low frequency words is a cost-benefit distinction: 'The cost is the time and effort to teach and learn the words. The benefit is the number of opportunities to use the words as represented by the frequency of the words' (1999:35). In other words, words that are more frequent should be learned first and the teaching of less frequent words should follow.

A further approach that goes beyond purely quantitative measures is the one by Laufer and Nation (1995) who argue that lexical originality (LO), lexical density (LD), lexical sophistication (LS), and lexical variation (LV) are amongst the most popular measures used for determining a learner's productive lexicon. Lexical originality is the percentage of words in a piece of writing that are used by one particular writer and no one else in the group (also called hapax legomena). Laufer and Nation report that this measure is quite unstable because it is defined by the group factor which means that if the group changes the index changes too (the performance value is relative to the group). This makes the measure unreliable as we can only get

information about a learner's performance in relation to the rest of the people in the group. Lexical density is the percentage of content words (nouns, adverbs, verbs, adjectives) in the text. Lexical words contain all the information, so if a text included more lexical words it should be considered denser. However, lexical density (LD) is influenced by the number of function words and this affects the validity of the measure. This means (according to Laufer and Nation) that it is not particularly certain that this index measures vocabulary as the lack of function words in a text could be the result of more subordinate clauses or ellipsis which are structural not lexical characteristics of a text. Lexical sophistication (as already discussed above) is the percentage of 'advanced' words in a text. The authors state that the weakness of this measure lies in the fact that 'advanced' is defined differently by different researchers, causing the measure to become unstable. Lexical variation (as previously mentioned) is another term for lexical diversity (for definition see Section on Lexical Diversity). This measure can be affected by differences in text length. Laufer and Nation also state that LV is dependent on the definition of a word. Laufer and Nation in 1995 introduced a new measure of lexical richness, the Lexical Frequency Profile (LFP). The VocabProfile (and its latest version, Range) software, which was developed by Laufer and Nation (1995), analyses the vocabulary of different texts, places them into four different lists (frequency layers), and gives an LFP (Lexical Frequency Profile) which shows the richness of each text. The LFP or Lexical Frequency Profile, which was proposed by Laufer (1994), shows the percentage (based on the total number of types in the text) of words used at different vocabulary frequency levels. The calculation is carried out by the VocabProfile computer programme (http://www.lextutor.ca/vp/eng/). This compares the text with the different vocabulary lists to see what percentage of the words are covered in the text. Laufer and Nation report that a word is defined as a word family in the programme (base form+ inflected and derived forms). The base word lists that are available for the programme are four. The first is based on the first thousand most frequently used words in English, the second includes the second thousand most frequently used words, and the third includes words that are not found in the two previous lists and are not used as frequently. The fourth layer includes words that are not found in the previous lists. The source for the first two lists is 'A General Service List of English Words' (West, 1953), and for the third 'The Academic Word List' (Coxhead, 1998; 2000). Therefore, words found in the first two lists will belong in the first two

thousand most frequent word families in English and words found in the third list will belong in the AWL (Academic Word List) and will be low frequency words (rare /infrequent words). The authors present their study using LFP in practice. The aim of the study was to establish the validity and reliability of LFP as a measure of lexical richness. Laufer and Nation expected to find the same LFP across different samples collected at the same stage of learning. In addition, they expected to see that at a higher level the lexis would be richer. If the LFP correlated with the Vocabulary Levels Test (Nation, 1983) it would show validity of the measure. The measure would also be considered valid if it distinguished between different levels of language proficiency. Using a sample of 65 foreign learners of English they collected two compositions, written from each subject during class time in one week. The length of each composition was 300-350 words and the topics of the compositions were of a general nature. The learners in the experiment were also given the active version of the Vocabulary Levels Test. All the compositions were entered into the computer, which only analysed the first 300 words of each composition. The researchers omitted any words that were incorrectly used from the count. The VocabProfile programme carried out the calculations. The results showed that the less proficient students used more of the first 1000 most frequent words. The less proficient also made use of the second 1000, but the most significant differences appear with the more sophisticated vocabulary, the UWL and the 'not in the list' words. 'These differences are in accordance with the concept of language proficiency which assumes that richer vocabulary is characteristic of better language knowledge. If the LFP has tapped these differences, this is evidence for its validity' (Laufer and Nation, 1995:316).

Laufer (1994) presents the weaknesses of the existing four accepted measures of lexical development and is in favour of adopting the LFP. According to Laufer, the LFP has many advantages over other measures of lexical richness. Unlike lexical originality, LFP does not change with the change of group; this makes the LFP a more objective tool. In addition, it is not dependent on syntax or text cohesiveness like lexical density. Furthermore, lexical sophistication only distinguishes between two types of words –frequent and sophisticated, while LFP provides a more detailed picture of the different types of words. Lastly, the LFP is free of subjective decisions regarding what a topic or thematic unit is. Thus it is more reliable than other less frequent measures. According to Read and Chapelle (2001), words that are used

incorrectly are excluded from the frequency analysis, making it a more reliable measure. 'The LFP included a procedure whereby content words which have clearly been used incorrectly by the learner are excluded from the frequency analysis' (Read and Chapelle, 2001:7).

Meara (2005) criticises the use of the LFP as a reliable tool for assessing L2 vocabulary because it is not as sensitive as claimed and cannot detect small changes in vocabulary size. It is only reliable when the groups compared have large differences in vocabulary size. LFP does not work well for learners that produce smaller essays because, according to Laufer and Nation (1995), two 300 word essays are needed to obtain stable vocabulary size estimates (Meara and Fitzpatrick, 2000). According to Crossley at al. (2011a), LFP can be less predictive, especially for shorter texts. Edward and Collins's state that 'the findings confirm that the ability of LFPs to distinguish between groups diminishes as vocabulary size increases. However, for fairly homogeneous groups, LFPs are able to provide a coarse but reasonable tool for vocabulary size estimation' (Edwards and Collins 2011:1). Laufer (2005) rejects Meara's criticism because he uses 'artificial data' and the bases of her rejection of 'artificial' data is that Meara got it from computer simulations and that Laufer does not think that they are valid for 'real-world' research on human learners. Although the computer simulations used by Meara to analyse LFP might give us some insights in the validity of LFP, they probably cannot cover all the complexity of real-life language learning and vocabulary testing. In my view, Laufer is right to reject Meara's findings to some extent because they are 'artificial'. However, Meara's criticism on the sensitivity of the LFP (to capture small or modest changes in vocabulary) should be taken seriously and it casts some doubts on the validity of LFP.

Meara and Fitzpatrick (2000) proposed Lex30, which is a word association task that stimulates vocabulary production and was designed to measure productive aspects of deep word knowledge. Word frequency data is used to measure the vocabulary produced (Fitzpatrick and Clenton, 2010). The test presents learners with a list of 30 stimulus words which they are required to respond to. One of the test's advantages is its practicality, because it is not time consuming and is easy to administer. Fitzpatrick and Clenton (2010) investigated the performance of Lex 30, which is a test of productive vocabulary, and suggested that is a valid test for vocabulary knowledge

because it produces consistent scores from learners over a short time period. Even though Lex30 seems to be a valid test of productive vocabulary it could not be used in my study due to the nature of my data- the test requires students to produce words based on a stimulus word, whereas I had essays to work with and analyse.

Meara and Bell (2001) outline the need to go beyond intrinsic measures of lexical variety (measures that are based on tokens and types) and develop extrinsic measures of lexical richness (i.e. measures that look at sophistication /frequency of words). Such measures would provide supplementary information about the tokens and types. The LFP (Laufer and Nation, 1995) discussed above is one of these measures, but, according to Meara and Bell, has some limitations. They propose P_Lex (2001), which explores the distribution of difficult words in a text. It produces a simple index that shows how likely the occurrence of these words is. Although P_Lex may seem similar to Laufer and Nation's (1995) Lexical Frequency Profile because they both look into the occurrence of infrequent words, there is a big difference between them as LFP just reports the percentage of these words whereas P_Lex uses a mathematical model. P_Lex is based on a computer programme that models the occurrence of rare words with a Poisson distribution. A Poisson distribution has a single parameter, 'Lambda', which can be used as a mark for the essay. A Poisson distribution gives the probability of obtaining exactly n successes in N trials (e.g. 4 rare words in 10 words). For this programme to work a basic word list is needed. Lambda values are easier to interpret and work with than LFP ratios (Meara and Bell, 2001). One of the reasons behind this is the fact that LFP gives you four different values, whereas P_Lex gives only one (lambda). Therefore, P_Lex may be easier to interpret. Lambda values normally range from 0 to 4.5, and the higher the figure, the higher the proportion of infrequent words. They are also less sensitive to text length than LFP scores, so P_Lex is more suitable for use with relatively short texts. Therefore the P_Lex methodology can be seen as reliable. Even though both the LFP and P_Lex use the same frequency list (Xue and Nation, 1984), P_Lex seems to have many advantages over LFP.

Guiraud Advanced is another measure of lexical sophistication proposed by Daller, Van Hout and Treffers-Daller (2003). This is wordlist based (which is similar to Meara's 'extrinsic' measures) and is calculated by dividing the advanced types

(words that are not in the basic lists) by the square root of the number of tokens (all tokens, not advanced tokens). Advanced TTR is a transformation of the TTR with the difference that the ratio is calculated by dividing the number of advanced types by the number of tokens. Daller et al. (2003) investigated existing measures of lexical richness in the spontaneous speech of bilinguals and proposed these two new measures: The Advanced TTR and Guiraud Advanced, which are suggested to have more advantages than traditional measures. They both lead to highly significant results which can be explained more clearly even with smaller samples. In their study the lexical richness of two groups of Turkish-German bilinguals was calculated. The new measures demonstrate the characteristics of the subjects better than the existing measures. 'The reason for the advantage of the advanced measures is the fact that they include additional information that is not available with purely quantitative measures' (Daller et al., 2003: 218). Even though both new measures -Advanced TTR and Guiraud Advanced- proved to be more powerful measures of lexical richness than the existing measures (TTR and Guiraud), because they had a wider scope and showed differences between the groups more clearly, the results were clearer when using Guiraud Advanced.

It must be mentioned here, to avoid confusion, that all previous measures, even though they are under the title measures of lexical sophistication (because they go beyond just looking at size but also look at frequency of words from various lists) are all measures of breadth not depth of vocabulary knowledge. All measures used in my research are measures of breadth, apart from one in Study 2 (collocations/formulaic count, which is an aspect of depth of vocabulary knowledge). Here follows a short discussion on measures claimed to investigate depth of vocabulary knowledge (see more on aspects of depth of knowledge in Chapter 4).

Wesche and Paribakht (1996) state that most research on second language vocabulary acquisition is based on measures of vocabulary size or 'breadth' measures, however, few researchers concentrate on 'depth' (terms of kinds of knowledge of specific words or degrees of such knowledge). Several vocabulary size measures are discussed and criticised in their work, and a new instrument called the Vocabulary Knowledge Scale (VKS) is proposed to enable researchers to assess levels of familiarity with given words. The VKS (Wesche and Paribakht, 1996) is an

instrument used for measuring depth of knowledge. The VKS is an instrument that efficiently captures certain stages in the initial development of knowledge of given words. This instrument elicits both self-perceived and demonstrated knowledge of specific words in written form by using a scale combining self-report and performance items. Here is an example of the scale from Paribakht and Wesche (1993):

*1:	I don't remember having seen this word before*

*2:	I have seen this word before but I don't know what it means*

*3:	I have seen this word before and I think it means _____ (synonym or translation)*

*4:	I know this word. It means _____ (synonym or translation)*

*5:	I can use this word in a sentence. e.g.: _____ (if you do this section, please also do section IV)*

Paribakht and Wesche's (1997) test is selective in nature as certain target words are selected to be the focus of the assessment (Read and Chapelle 2001). This instrument was used in a pilot study (conducted by Paribakht and Wesche) whose results revealed significant intra-group gains. VKS also proved sensitive to inter-group differences in content vocabulary gains. Paribakht and Wesche report that minor changes and clarifications were made to improve the instrument's precision for further studies.

However, the authors state that the purpose of VKS is to capture the different initial stages of word learning and not to estimate general vocabulary knowledge. In addition, it does not reveal anything about understanding different meanings of the same word or different aspects of word knowledge. It is also unsuitable for large samples because it requires hand-scoring. Paribakht and Wesche state that one of the main advantages of VKS is the fact that it elicits students' perceived knowledge of vocabulary items and also allows verification with demonstrated knowledge. Demonstrated knowledge here is showed by the learners ability (if they choose statement number 3, 4 or/and 5) to produce a synonym or use the word in a sentence. However, from my point of view, one of the main disadvantages of the VKS is the

fact that learners need to assess their vocabulary knowledge (choosing between the five statements of the scale). Therefore, they can always underestimate or overestimate what they know or how they know it resulting in non-representative results (as Read suggested in 1993- it is not appropriate to rely on self-report). For example, they may use the word in a sentence that does not show the true meaning of a word. For example, if we take the word 'beautiful' the learner could respond to VKS Statement 5 with this sentence: *I don't really know what beautiful means.* This according to the VKS scoring criteria is a grammatically correct sentence but does not show the meaning of the word 'beautiful' so there would be problems with assigning a score to this sentence. In addition, another disadvantage is that it is very time-consuming. The VKS is also criticised by Henriksen (1999) as to whether it really measures depth of knowledge.

Read (1993) emphasised the need for suitable instruments with which to measure vocabulary acquisition and reports the investigation of a new test format that will also test how well particular words are known, not just if a word is known. The test (Word Associates Format, WAF also known as Word Associates Test, WAT) was designed for measuring vocabulary acquisition in students learning English for academic purposes at university level. He states that they wanted to develop a test format that would ask for a simple response, but would have a large coverage of words and at the same time would test depth of vocabulary. The concept of word association was initially used due to the fact that there is extensive literature on word associations in L1 (Deese, 1965; Clark, 1970; Postman and Keppel, 1970) and L2 users and learners (Meara, 1980:234-39; Meara, 1983). Even though Meara had decided that word-association tasks were not satisfactory for testing learners' vocabulary knowledge, he suggested to Read to create a task in which learners would choose answers instead of giving their own (learners are presented with a stimulus word). Stimulus words were selected from the University Word List (UWL) Here is an example of a word associates item (Read 1993:359):

edit

arithmetic   film   pole   publishing
revise   risk   surface   text

The evidence from Read's analyses show that this word associates test is reliable and can be used to measure knowledge of academic vocabulary as represented by the University Word List (Read 1993:368). Read conducted an item analysis to prove the test was reliable and also had a 'verbal report' from eight students explaining the deciding factors for their choices. However, the results can be affected by the test-takers willingness to guess the correct answers because in most cases people that were willing to guess the answers were successful. The items of this test are heterogeneous in structure in a variety of ways, and this characteristic reflects the actual variety of words in the language and is appropriate for a test that was designed to measure the quality of vocabulary learning in a university EAP course. However, the test was rather complex to analyse and the analysis showed that there was a high variability in the patterns of responses to individual items. To help improve the Word Associates Format and allow it to reach its full potential as a research tool, it is necessary to develop tests that focus on more homogeneous subsets of vocabulary items. To obtain more conclusive results it is essential to have a larger dataset, because Read's study was limited by the relatively small number of test-takers. Schmitt, Ching and Garras (2011) criticise the WAF (Word Associates Format), which is a test of depth, as they state that the WAF can sometimes underestimate or overestimate word knowledge. Nevertheless, the test is used as a measure of depth of knowledge in various studies (Ehsanzadeh, 2012). My main critical comment regarding the use of the Word Associates Format is the same as the VKS, therefore I would not be eager to use it due to its reliance on self-report. In addition, word associations are very difficult because everybody has other/different associations.

2.5 METHODOLOGICAL PROBLEMS WHEN MEASURING VOCABULARY

2.5.1 Problems with definitions

There are various problems regarding the existing vocabulary measures. The first which was discussed in a previous section is the problematic definition of lexical diversity. According to G Yu (2009) there are many terms that are used by various researchers interchangeably (terminological challenges). G Yu (2009:238) also states that: 'Further complications arise when the same term was conceptualized and quantified differently in different studies. Indeed, different conceptualisations and

quantifications of lexical *diversity* make it difficult to compare and synthesize the findings of these studies although they used the same term- *lexical diversity.*'
G Yu (2009) suggests that even though there are many empirical studies in applied linguistics that have measured lexical diversity (as an indicator of writing and speaking performance), it is extremely hard to compare their findings due to the fact that they have used different conceptualisations and quantifications of lexical diversity. The many different names and operationalisations of lexical diversity make it hard to compare research findings.

2.5.2 How do we count words?

Apart from the definition problems and the fact that sample size can affect the results of some quantitative measures, a problem lies in the fact that it is hard to operationalise the construct of vocabulary (Schoonen, 2001; Read, 2000). It is difficult to discuss or analyse quantitative aspects of vocabulary because it includes counting/numbers; in order to do this the researcher first needs to decide on a definition of what a word is and what to count as a word (Nation, 2001). Vocabulary size has been hard to measure due to serious methodological problems revolving around what we count as a word or how we test whether a word is known (Nation and Waring, 1997). Since 1942 (Cronbach) and for many years to follow researchers (Richards 1976; Nation 1990) have been struggling to decide what constitutes a word. The concept of 'word' is very unclear (Bogaards, 2000). According to Carter (2000), an orthographic definition is very simple: a word is any chain of letters which has on each side either a space or a punctuation mark. This definition is used for practical and testing reasons. This definition is quite simple and easy to use when counting words, but there are certain problems involved when one actually starts counting. For example, even if we decide on the orthographic definition, how do we count words such as *run*, *runs*, *running* and *ran*? Should they be counted as four separate words (Carter, 2000)? As already mentioned in a previous section, it is not sufficient to view vocabulary only as single words as vocabulary knowledge is much more complicated and entails more than just knowing a dictionary meaning (Fatahipour, 2012). The main proposition by Carter was to count lexemes (the base forms of words as they are found in dictionaries.). Another issue raised is how to measure compound words such as post box (Carter, 2000). A definition given by Bloomfield (1933, cited in

Carter, 2000) states that a word makes sense on its own if it is used as an answer to a question, a statement or exclamation. Read and Chapelle (2001) refer to the 'ill-defined nature of vocabulary as a construct' (2001:1). According to them, different researchers approach vocabulary from different angles. It is problematic to attempt to define what to count as a word (Coxhead, 2000). Gardner (2007) is very critical of the construct of a word (what constitutes a word) for research and pedagogical purposes. Word families, multiple meanings and multiword items are all aspects that affect the validity of the construct of word. 'Words may seem like simple entities but they are not. Their surface simplicity belies a deeper complexity' (Pearson et al., 2007). In Beglar and Hunt's (1999) study, 'a *word* is defined to mean a base word plus all of its inflections and derivatives' (Laufer, 1992, 1997; Nation, 1990; Read, 1998). Thus, the base form *buy,* plus its inflections and derivatives- *buys, bought, buying, buyer* and *buyers*- constitute a word often termed *lexeme*- (Beglar and Hunt, 1999:133). One could argue of course that derivational morphology which changes the word class creates new words, therefore *buy* and *buyers* are two different types. Beglar and Hunt (1999) argue that if a learner knows the base form there is as yet no evidence that they will know all the derived and inflected forms of that word.

Various researchers have tried to measure different aspects of vocabulary knowledge (Schmitt and Meara, 1997; Laufer, 1998; Read, 1998). It should be pointed out that according to Read (2000), one of the main problems with measuring vocabulary size is the fact that some researchers focus on counting word forms and others focus on counting word families. Nation (2008) questions whether we should count different forms of the same word as different words or not. Word forms are different forms of a word, such as *wait*, *waits*, *waited* and *waiting*, which are also known as lemma. Words with different morphology could sometimes be strongly related to be accounted as one single item. Therefore, to overcome this problem most lists consist of word families (West, 1953; Xue and Nation, 1984) - see Glossary of Terms and Abbreviations for a description of 'word family'. Researchers who carry out studies that involve counting words lemmatise the tokens, therefore the base word (in this case the word *wait*) will only be counted once since the inflected forms will not be counted. A word family is a group of words that share the same basic meaning- they share a common base (Read, 2000). Raw or lemmatised data could give different results of lexical diversity according to Richards and Malvern (2007).

2.5.3 What about Multiword Units (MWU)?

Another key issue is how to deal with multiword units (MWU), and what to consider as a MWU (Pawley and Syder, 1983). These units include collocations, idioms, formulaic sequences etc. This will be pursued in Chapter 4 (for a detailed discussion see Chapter 4). It is extremely difficult to try and count these MWUs (Nation, 2008). However, they are very important in defining someone's vocabulary size (Nation, 2001). This issue was also raised by Carter (2000) and Read (2000), who were concerned about the way idioms such as *kick the bucket*, should be treated if they are to be counted as three separate words. If we follow Bloomfield's definition they should be counted as three different words, but if we do this they lose the meaning they have as a multi-word unit. Carter (2000) is very critical of the vocabulary measures because he argues that researchers cannot measure someone's vocabulary until certain problems are overcome and definitions are agreed upon.

To summarise, it is obviously difficult how to define a word, and researchers need to be clear on the definition they would like to use when they engage in research involving vocabulary measurement. I do not use a single definition in my work (I use different definitions in different studies) as I believe that none of the definitions above completely cover the concept of what a word is and how it should be counted. The main definition used for my research is the Orthographic Definition by Carter (2000), which treats words as any chain of letters with either a space or a punctuation mark on their side. For Study 1 I do not count lexemes (the base form of words) because when a leaner knows a word it does not necessarily mean that they know all its derivatives. This is supported by literature - see Chapter 3, Section 3.3.2. For the purposes of Study 2 however, I count chunks of language (formulaic sequences). Thus, it can be assumed that a combination of different definitions and aspects of lexical knowledge are used in this thesis.

2.5.4 Small (and unrealistic) amount of data

Daller et al. (2007) state that researchers usually use a small amount of language for their analyses. This can be quite problematic due to the fact that a small amount of language may not be representative of what a learner knows, so a single short piece of speech may tell us little about the amount of productive word knowledge a learner

has. According to G Yu (2009:238), 'the lexical diversity of a product is only one static manifestation of the producers' lexical diversity which may well be dynamic in nature'. Moreover, Möbarg (1997:212) argues that:

> 'testing vocabulary status in production immediately poses problems, however, it is a well-established fact that any given text only employs a tiny fraction of the author's full vocabulary and, furthermore, that the text will determine, i.e. delimit, the scope and choice of vocabulary used'.

However, our only option is to use a small amount of language because we cannot monitor someone's everyday speech from day to day -we cannot be present every single moment our subjects produce language of any kind.

2.5.6 Choice (and influence of topic) and setting

Furthermore, the choice of topic can have a different effect when measuring learners' vocabulary. G Yu (2009:254) states that 'compositions on impersonal topics had significantly higher lexical diversity than personal topics. Higher lexical diversity was achieved when candidates were highly familiar with the topic'. Research by Brown (2003:53) revealed that, when rating oral IELTS interviews, examiners commented on the adequacy of candidates' vocabulary for the type of topic (describing it with terms such as *familiar, unfamiliar, professional* etc.). Finally, Cook (2008) states that it is not enough to just count words in laboratory settings. Tests need to be developed in order to find out if people can use the words or can remember them.

2.6 LANGUAGE PROFICIENCY

2.6.1 Definition

'Proficiency in a second language is one of the most fundamental concepts in Applied Linguistics, and accordingly its character is the subject of ongoing and intense debate' (Iwashita et al., 2008:24). Language proficiency is very difficult to define. Individuals give different answers when asked to define language proficiency. One of the broadest terms, suggested by Blue, Milton and Saville (2000), is the amount of language a person is acquainted with. There is also ambiguity in the use of the term 'proficient', and it can be used interchangeably with other words -for example,

competent, good, fluent (Iwashita et al., 2008). One of the main components of academic proficiency is academic language proficiency, the other is knowledge of academic content (Krashen, 2011). Proficiency levels can be distinguished by various features of test-takers discourse under analysis, such as vocabulary (token and type), grammatical accuracy, grammatical complexity, pronunciation and fluency as defined by Iwashita et al. (2008). Crossley et al. (2011b:182) also state that '…lexical proficiency is an important element of language proficiency and fluency, especially for second language (L2) learners'. In addition, according to Laufer et al. (2004), vocabulary size is linked with general language proficiency.

2.6.2 Lexical Richness and Proficiency Ratings

There is an on-going discussion regarding the role of lexical richness within the construct of foreign language proficiency (Daller et al., 2007). Could the use of infrequent words be an indicator of language proficiency? The use of infrequent words seems to reflect a greater vocabulary size and sophistication (Wesche and Paribakht, 1996). The use of certain function words can also indicate the proficiency level of learners (Morris and Tremblay, 2002).

Laufer and Nation (1995) suggested that a richer vocabulary is an indicator of a better understanding of language, and wanted to discover whether there would be a significant difference between the LFP's of learners of different language proficiency levels. Their results showed that the less proficient students were using more of the first 1000 most frequent words, and therefore their hypothesis was confirmed (Laufer and Nation, 1995).

The results of Morris and Cobb's study (2004), who used vocabulary profiles as predictors of the academic performance of TESL (Teaching English as a Second Language) trainees, showed that the more words (tokens) produced by a learner the higher the level they achieved. This was also the case with a wider range of words-types (Iwashita et al., 2008). Various studies showed that different aspects of lexical proficiency can be predictive of L2 production (McNamara, Crossley and McCarthy, 2010; Crossley et al., 2011a). The results of a study by Iwashita et al. (2008) show that the features of vocabulary and fluency (as individual detailed features of spoken

language produced by test takers) have the strongest correlation with levels of performance (speaking proficiency). Adam's (1980) study, which examined the relationship of 5 different components (accent, comprehension, vocabulary, fluency, and grammar) of the Foreign Service Institute (FSI) Oral Interview Test of Speaking and the global speaking scores, showed that vocabulary and grammar were the main components (factors) that distinguished different levels of proficiency. Higgs and Clifford (1982) proposed the Relative Contribution Model (RCM) due to their suggestion that 'different factors contribute differently to overall language proficiency'. This model describes 'rater perceptions of the relative role of each of five component factors making up global proficiency -i.e. vocabulary, grammar, pronunciation, fluency and sociolinguistics (Iwashita et al., 2008:26). Teachers in Higgs and Clifford's study (1982), as results show, thought that vocabulary and pronunciation mattered most at lower levels. This changes at higher levels where all four components- apart from sociolinguistics- seem to have equal weight (Iwashita et al., 2008). In Hawkey and Barker's study (2004) it was found that at higher IELTS proficiency levels essays were longer and employed with a broader vocabulary. After the use of a standardised version (compute every *n* words rather than once *n* for the whole text-the default is every 1000 words) of the type- token ratio, which allows the comparison of texts of different lengths, it was confirmed that vocabulary range increases as proficiency levels increase. Therefore, 'range of vocabulary is thus possibly a feature distinguishing proficiency levels' (Hawkey and Barker, 2004:143). G Yu (2009) states that lexical diversity is a predictor of general language proficiency as his results revealed that D correlated positively and significantly with language proficiency.

Daller and Xue (2007) also investigated different measures of lexical richness in order to find which measure is the best to use for measuring oral proficiency. They asked participants to describe two picture stories; their descriptions were recorded and transcribed into CHAT. Then the participants' lexical richness was calculated by using various measures such as D, P_Lex, LFP, TTR, Guiraud Index and Advanced Guiraud (see Glossary of Terms and Abbreviations for explanations of these measures). The results showed that the most appropriate measures for oral data were Guiraud's Index and D which yield lower and highly significant p-values when groups are compared. In addition, even though Advanced Guiraud and LFP showed

the differences between the groups, they were not as suitable as the previous measures for the given context. According to the researchers this may be due to the fact that these word-list based measures were not developed on the basis of everyday spoken language. The only measure which seemed to be invalid measure for oral data (spontaneous speech data) was TTR which did not produce a significant p-value. As previously mentioned, Fatahipour's (2012) study showed that lexical richness measures can be used to partially address the construct of language proficiency. There was a correlation (but not a strong one) between language ability and lexical richness measures.

Tidball and Treffers- Daller (2007) state that the measures D, Guiraud Index and Advanced Guiraud are all valid for measuring lexical aspects of language proficiency (which is also supported by Daller and Xue, 2007). The results from their study embrace the suggestion by Malvern and Richards that researchers should not only use one single measure in research but rather a combination of all of them which could lead us to a better understanding of people's vocabulary knowledge.

2.6.3 Lexical knowledge and reading ability

Nation (2001) states that there is a close relationship between lexical knowledge and reading comprehension. This is also supported by Shen (2008) who argues about the existence of such a relationship and explains that their connection is complex and dynamic. Vocabulary size is found to be directly linked to reading comprehension (Stahl, 1999). In their study, Albrechtsen, Haastrup, and Henriksen (2008) found a significant correlation between L2 vocabulary size and L2 reading ability. Laufer (1992) conducted research on how L2 lexical knowledge interacts with the reader's general academic ability. She wanted to investigate how L2 proficiency affects L2 reading. Her results showed that lexical richness in L2 is a better predictor of reading in L2 than a learner's general academic ability (including the reading ability in L1). She suggests that lower proficiency learners could improve their L2 reading skills by improving their vocabulary knowledge. 'Laufer concludes that a vocabulary of less than 3000 words is a more significant factor in limiting English reading comprehension for academic purposes than learners' general academic ability, including L1 reading skills' (Beglar and Hunt, 1999:134).

Her results reinforce the suggestions of other researchers (e.g. Kelly, 1989), which state that:

> 'Vocabulary constitutes the single largest obstacle to advancement and a massive vocabulary instruction programme is of the utmost importance in the teaching of a foreign language. If a good knowledge of foreign vocabulary can compensate for lower general academic ability, then even learners of mediocre ability can improve considerably in their L2 reading once they have raised their lexical level' (Laufer, 1992:101).

Scores on depth of vocabulary were also found to be good predictors of reading comprehension levels (Akbarian, 2010).

Pearson et al. (2007) state that vocabulary is an important factor in text comprehension. This idea is supported by Hirsh and Nation, who suggest that readers need to be familiar with 95 per cent of the words in a text to comprehend and understand its main points and use their L1 reading skills to read in a second language (Hirsch and Nation, 1992). Besides Hirsh and Nation's study there are various others that support the idea of vocabulary being an important aspect of text comprehension. However, there seems to be a disagreement in the percentage of words that need to be known for text comprehension. Qian and Schedl suggest 'that vocabulary knowledge is instrumental in reading comprehension…' (2004:28), and Schmitt, Wun Ching and Grabe (2011) suggest that the percentage of words needed for reading comprehension is estimated to be around 98%, as suggested by Hu and Nation (2000). Furthermore, Nation (1990) suggests that knowing 3000 words is enough to understand 95% of general texts. In academic texts, 3000 words covers 88% of a text. 'Qian (1998; 1999; 2000; 2002) has found that in reading comprehension both depth and breadth of vocabulary knowledge play important roles, and that two aspects of depth of vocabulary knowledge- namely, meaning, which includes synonymy and polysemy, and collocation- are important variables' (cited in Qian and Schedl, 2004:30). These findings seem to also apply to spoken discourse as a study by Adolphs and Schmitt (2003) showed that 'around 5,000 individual words were required to achieve about a 96 per cent coverage figure. These results suggest that more vocabulary is necessary in order to engage in everyday spoken discourse than was previously thought. The implication is that a greater emphasis on vocabulary development is necessary as part of oral/aural improvement' (2003:425). The latter study does not concern reading ability but spoken discourse. However, it was

included in the discussion because it is quite relevant, as it shows the importance of vocabulary knowledge in comprehension in general. Even though researchers seem to conclude different findings, there seems to be a consensus regarding the fact that vocabulary knowledge is linked with comprehension.

2.6.4 Lexical knowledge and school success

According to Verhallen and Schoonen (1998), lexical knowledge is an important predictor for school success. The results from their study showed that bilingual children are disadvantaged at school because not only do they know fewer words in their L2 than in their L1, they also have a shallower knowledge of the L2 words they seem to have acquired (the meaning allocation to these words is poorer and less paradigmatic).

2.7 SECOND LANGUAGE (L2) TESTING AND SCORING

2.7.1 Language testing

Language testing research has evolved (like every research brand) throughout the years. There were two main assumptions before the 1980s regarding the dimensionality of language proficiency and measurement: it was assumed that language proficiency was one-dimensional (uni-dimensionality), and quantitative research, using statistical methods, was the norm (Bachman, 2000). These assumptions developed in the 1980s and it was believed that proficiency was a multi-trait construct and the need for communicative language tests arose (Alderson, 1981). In the 1990s the research broadened further with the expansion of research methodologies, the development of authentic tests, concerns about ethics and aspects that could affect performance were investigated. Nowadays, the concept of language testing is one of main areas of applied linguistics and applied linguistic research. According to McNamara (2011), the problem with language testing is the fact that researchers approach it from a single perspective: either by statistics and measurement (the 'testing' side), or language linguistics (the 'language' side), and not both as should be the case (McNamara, 2011:435). This means that testing should be more spherical (be approached from various angles/perspectives) and constantly updated by theories of psychometrics and also theories of language use. According

to Alderson (2005), diagnostic tests have certain characteristics that make them distinguishable from placement or proficiency tests, and the most significant of these is the fact that these tests should be used to detect a learner's strengths and weaknesses and should be specific rather than global.

2.7.2 Testing vocabulary

Testing vocabulary plays a major part in the role of language testing in general. As previously mentioned, there are tests of vocabulary breadth and depth. The importance of vocabulary breadth tests is highlighted by the following statement by Laufer et al.:

> 'Depth tests tend to test only a small number of items, their value lies mainly in enabling us to research specific items targeted for investigation amongst specific research participants. Size tests, on the other hand, consist of larger samples of words from different word frequency levels, which, when chosen randomly, represent the entire vocabulary at these levels.' (Laufer et al., 2004:208)

Bogaards (2000) suggests that when testing L2 vocabulary knowledge there are many aspects that should be considered and tested, such as meaning, morphology, syntax, collocations etc. 'Testing vocabulary knowledge in a second or foreign language is not as straightforward an affair as is sometimes thought' (2000:490). An issue regarding vocabulary testing is whether to test words in context or in isolation. Should vocabulary be tested in context? Schmitt (1999) states that some learners may recognise words in context, but not when isolated, raising the question of what is actually tested: vocabulary knowledge or inferencing skills.

2.7.3 Influence of context in testing vocabulary

When testing vocabulary it is very important for any researcher to think about the influence context may have on their testing. Some examples of tests that analyse individual words and do not take context into account are the Yes/No Format (Meara and Buxton, 1987) and Levels Test (Nation, 1990).

Read (2000) contributed a detailed description of the difference between two types of tests who names discrete vocabulary tests and comprehensive (embedded) ones.

Discrete tests are tests in which words are tested as single items (are isolated from any context- context does not play a role in the assessment), basically tests that focus on selected target words. Comprehensive (or embedded) tests not only test vocabulary items in isolation but 'vocabulary is embedded as one component of the measurement of a larger construct, such as communicative competence in speaking, academic writing ability or listening comprehension' (Read 200:188). There is one thing that we need to consider though: according to Read not all comprehensive measures are embedded ones since some researchers use them on a discrete basis. This means that even though they may have a large sample of text produced by a learner they isolate vocabulary and try to measure it and are not interested in assessing any other abilities. This is the approach I adopt in my studies too because even though vocabulary is presented in context (in the form of an essay) I am not really testing words in context but in isolation. Of course context plays an important role in testing vocabulary so a short discussion on the influence of context in testing vocabulary will follow but it is not in the scope of this study to examine this any further. Read (2000) justly wonders whether vocabulary can be separated and tested on its own and not as part of language proficiency in general (see discussion on language proficiency in Section 2.6). He argues that one of the disadvantages of testing a word in isolation is the fact that the word could have multiple meanings and there would be no clue as to which word the researcher is attempting to assess. However, if we present words in a sentence, learners could guess or infer the meaning from surrounding words. There is not much research on the role of context in vocabulary assessment but it is generally agreed (and followed in this study too) that is it best to present vocabulary in context.

2.7.4 Test Reliability and Test Validity

Researchers need to be extra careful when designing and running a test regarding the test's reliability and validity. Test reliability refers to the notion of how accurately a test measures what it is supposed to measure. Therefore, if a test is reliable it would mean that if you run/repeat the test several times you would get the same results. Test validity refers to 'the extent to which a test measures what it is supposed to measure' (Daller et al., 2007:16). We always need to be sure about the concept we are testing. Tests that are used to measure vocabulary need to be tested for their reliability and validity. When testing vocabulary it is very hard to know if what is tested is actually

what is supposed to be tested. Therefore, other issues arise such as content validity and construct validity. Content validity refers to whether a test has the appropriate content, and construct validity checks whether what is measured is the construct that is supposed to be tested. There is also the concept of convergent validity (see Glossary). Nation and Beglar (2007) argue about several issues that can threaten the validity of tests of vocabulary knowledge, such as the candidates' attitude towards the test or how willing they are to participate. In addition, they argue about the appropriateness of frequency data and suggest that tests that are based on L1 frequency lists could be less useful in second language conditions. This could be a possible reason for my Study's (Study 2) unexpected results- because the list used for the analysis in Study 2 is based on L1. The other issue is that of what each researcher is actually counting (which is discussed in a previous section). It makes a difference and the studies incomparable if one researcher counts lemmas and another one does not. Lastly, they suggest that the language of instruction could affect the test's validity. All these issues need to be taken into account when choosing any tests or measures for vocabulary testing or assessment.

2.7.5 Rating scales- holistic scores

In the present study raters were asked to provide an IELTS holistic (overall) rating for the essays (see Chapter 3- Methodology Section). Therefore it would be appropriate to present some issues regarding this type of rating scales.

'Holistic scoring is widely used in second language (L2) writing assessment' (Barkaoui, 2010:516). It has been broadly used in various large-scale writing assessments, such as the computer- based TOEFL (Lee, Gentile and Kantor, 2009). Holistic scoring is also called global or impressionistic scoring, according to Lee, Gentile and Kantor (2009). Recently, holistic scoring has started being used, especially in automated essay scoring and evaluation (Lee et al., 2009).

However, even though this type of scoring is widely used, there are various issues and limitations regarding the use of this holistic (or global) scoring.

> 'In particular it allows raters to include evaluation criteria not listed in the scale and to use personal judgement to determine how important a

specific [criterion] is to the overall score, thus resulting in raters moving away from the criteria originally designed to define what is being assessed. This can reduce score consistency across and within raters and, ultimately, change the meaning of the scores' (Goulden, 1994:74).

Connor-Linton (1995) also suggests that unless we investigate further with a more qualitative analysis (maybe with a think-aloud protocol) what the raters actually rate, we cannot be sure of what it is that the given rating represents. It needs to be considered that when using holistic rating the researcher can never be fully aware of what it is that is being assessed. Even though we may get the same ratings from two different raters, it does not necessarily mean that the same score had been awarded for the same reasons. Douglas and Selinker (1992, 1993) argue that different reasons may drive raters to arrive at the same ratings even if they use the same scoring rubrics. Connor-Linton (1995) also states that holistic ratings are not ideal for assessment/rating.

Other researchers also state that it is not very useful in assessment, and the use of this rating in writing research fails as a qualitative research tool. Hamp-Lyons (1995:760) states that 'a holistic scoring system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter'. Weigle (2002) criticises the use of rating scales in performance assessment by stating that they are not specific enough, leading the raters to a holistic marking. According to Weigle, this type of scoring is not suitable for picking up learners' particular weaknesses or strengths. In the case of second language learners this can be a major problem as learners may still be in the process of developing/acquiring writing skills and may produce uneven profiles for different aspects of writing. Holistic rating is not ideal for generating diagnostic feedback as it is multi-trait scoring, which is not used widely for very important reasons. These reasons are, firstly the cost, and secondly, the fact that the different traits are often interrelated and correlate highly among themselves and holistic scores (Lee et al., 2009). Therefore, holistic rating is reliable for identifying proficiency (levels), but cannot be used for identifying specific areas of weakness (Erling and Richardson, 2010).

Knoch's (2009) study compared two rating scales, for EAP (English for Academic Purposes) writing, one of which is more detailed than the other because it was empirically developed (with detailed level descriptors). The results showed that the

rater preferred the more detailed scale because they could differentiate the various aspects of writing. In addition, the raters' reliability was higher when the latter descriptors were used. In a comparison of holistic and analytic scales it should be noted that analytic scales have higher reliability and validity but are expensive and time consuming. However, they are more effective (Knoch, 2009).

Rating scales have been criticised for using 'impressionistic terminology which is open to subjective interpretations' (Knoch, 2009:278; Brindley, 1998). Mickan (2003, cited in Knoch, 2009) stated that band levels do not provide specific descriptions for each level, but rather a relativistic wording to differentiate between levels. Knoch's study showed that 'a rating scale with descriptors based on discourse-analytic measures is more valid and useful for diagnostic writing assessment purposes' (2009:301).

In the IELTS rating procedure all the scores from each section are averaged and rounded to produce an overall band score. The results are reported as whole and half bands. There is a problem of inconsistency in ratings and Mickan (2003) suggested that even though raters should use analytic scales they tend to rate the essays as whole (give a holistic rating) than distinguishing between different parts.

Lastly, another issue regarding holistic ratings is the existence of a halo effect. Similar ratings regarding the lexical and holistic ratings in this study may suggest the existence of a halo effect. 'Holistic type rating often results in a halo effect where a rater awards the same score for a number of categories on the scale' (Knoch, 2009:294). Knoch (2009) suggests that the existence of a halo effect is usually present when raters encounter problems in the rating process.

Despite the fact that the existence of a halo effect could always be a potential problem or hindrance to any study, there were only holistic ratings produced from the raters in all of my studies. This does not undermine or compromise the study because previous research (Zughoul and Osman Kambal, 1983) – which compared the holistic and analytic methods- showed that the inter-rater reliability was higher in the holistic rather than in the analytic rating.

2.8 IELTS

2.8.1 The IELTS test components

IELTS stands for 'International English Language Testing System' and is designed to provide students with evidence of their English proficiency (Blue et al., 2000). It is a very popular test which is recognised worldwide. It is designed for people who intend to study or work in an English speaking country. It measures the candidates' abilities in English across all four language skills (writing, speaking, listening and reading). Candidates can choose to take either the General training test or the Academic test. The first test prepares people to live in an English speaking country and be able to communicate and work. The second test prepares candidates for academic study in an English speaking university (British Council, http://www.britishcouncil.org).

For the purposes of the present study only the Academic test, which measures proficiency, is of interest. Details for each section of the Academic test can be found here:

2.8.2 The Academic Reading Test

This test is divided into 3 sections, each with 40 questions based on 1 reading text per section. The length of the Academic reading test is between 2000 and 2750 words. A question paper and an answer paper are given to all candidates. The candidates are allowed to write on the question paper but they cannot remove it from the test room after the end of the test. Candidates must put all answers onto the reading answer sheet before the end of the hour. There is no extra time allocated after the 1 hour set for the academic reading test for the transfer of answers to the answer paper.

Various question types are used for the tests and are usually selected from the following list:

- Multiple choice
- Short answer
- Sentence completion
- Notes/summary/diagram/flow chart completion
- Choosing from a heading bank to identify paragraphs or parts of the text

- Identification of writers opinions/ideas - yes/no/not given
- Identification of information in the text - yes/no/not given OR true/false/not given
- Classification
- Matching lists or phrases

Texts are taken from a variety of sources such as magazines, journals, books and newspapers. Texts do not require specialist knowledge of the subject. All reading passage topics are of general academic interest. At least one text contains a logical argument and one of the texts may include a diagram, graph or illustration. If there are any words or terms of a specialist technical nature, which candidates would not be expected to know, then a short glossary is provided.

2.8.3 The Academic Writing Test

The Academic Writing Test lasts one hour. Candidates are required to perform 2 tasks.

In Academic Writing Task 1, candidates are asked to describe in their own words factual information given to the candidate in pictorial form(s). The pictorial form(s) are usually a line graph, a bar chart, a pie chart, a table or a picture describing a process. Sometimes there could be a combination of these input forms. Candidates are required to write a minimum of 150 words.

In the Academic Writing Task 2, candidates are asked to write an essay on a general academic topic. Candidates must write a minimum of 250 words.

2.8.4 The Academic Speaking Test

The IELTS Academic Speaking Test is the same for both the Academic and General Training modules. The test is conducted by one examiner and one candidate and the conversation is recorded. The Academic Speaking Test is divided into 3 sections.

Section 1 The Academic Speaking Test Section 1 starts with some general introductory questions (*How are you today*? etc.). Then the candidates must answer questions relating to personal information, similar to the type of questions one would

ask when meeting someone for the first time. Finally, the examiner poses a series of questions on two topics of general interest (4 - 5 minutes).

Section 2 In the Academic Speaking Test Section 2, only the candidate speaks (it is a monologue by the candidate). The candidate receives a card from the examiner, which provides, a subject and a few guiding questions. The student then has to talk for 1 to 2 minutes on that specific subject without being interrupted by the examiner. The examiner determines the exact length of time. The students have an optional one minute to prepare for their talk and are given a piece of paper and a pencil with which to make brief notes. After the candidate's talk the examiner asks one or two brief questions to finish off the section (3 - 4 minutes).

Section 3 In the Academic Speaking Test Section 3, some more questions, generally related to the subject spoken about in Section 2, are asked by the examiner. These questions require some critical analysis on the part of the candidate and are usually more demanding (4 - 5 minutes).

2.8.5 The Academic Listening Test

The IELTS Academic Listening Test is the same for both the Academic and the General training modules. The candidates listen to a tape and then answer a series of questions. The candidates have to listen very carefully because the tape is played only once. The Academic Listening Test is divided into four sections, with 10 questions in each section (a total of 40 questions) and lasts for about 30 minutes. Candidates have an extra 10 minutes at the end of the test to transfer their answers to the answer sheet.

A variety of question types are used in the Academic Listening Test, usually taken from the following list:

- Multiple choice
- Short answer
- Sentence completion
- Notes/diagram/flow chart completion

(IELTS Help Now, http://www.ieltshelpnow.com)

The scoring system used is a distinctive nine point system. Each candidate receives scores for each language skill and an Overall Band Score on a scale from Non User (1) to Expert User (9) (McGovern and Walsh, 2007). As already mentioned in the introductory chapter, there has been an increase of IELTS test takers from around the world. IELTS is a test of great importance therefore justifying my decision to use it in my research to create a predictive model which, long term, could have financial benefits for test takers.

2.8.6 IELTS and vocabulary knowledge

There are three main studies regarding the relationship between IELTS scores and vocabulary/lexical knowledge. The first study is a study by Read and Nation in 2002 who examined vocabulary use in the IELTS Speaking Test. They decided to investigate vocabulary because Lexical Resource is one of the main criteria examiners need to rate for the IELTS Speaking Test. The researchers looked into the vocabulary items used by candidates, and their lexical diversity and lexical sophistication were measured. By conducting a more qualitative analysis the researchers also looked into the use of formulaic language by the candidates. Transcriptions of 88 IELTS Speaking tests were used for the calculations, and the results showed that the measures of lexical diversity (vocabulary size) 'did not offer a reliable basis for distinguishing oral proficiency levels' (Read and Nation, 2002:207). Therefore, the scores (band levels) of the IELTS Speaking Test could not be predicted by measures of lexical diversity. The qualitative analysis showed that higher band candidates used more formulaic language in their speech, but did not use as many low-frequency words. This is one of the studies that influenced my methodology design. My method was similar to theirs but not my findings. They measured lexical density (proportion of content words in a text) which I did not and is one of the differences between the two studies. In terms of measuring lexical variation and lexical sophistication similar calculations were carried out. They calculated D (lexical variation) and used P_Lex for lexical sophistication which was also used for my study. However, I also used other measures for lexical variation

(TTR and Guiraud) and lexical sophistication (Guiraud Advanced) -see Chapter 6, Section 6.3 for similarities or differences in our findings.

Mayor, Hewings, North, Swann and Coffin in 2002 investigated differences between high and low-scoring scripts (writing) of IELTS Academic Writing Task 2 using data from Chinese and Greek L1 candidates. They looked at error analysis, sentence structure, argument structure -at the sentence level and at the discourse level- and tenor and interpersonal meaning. They also conducted an exploratory qualitative analysis which involved only a small number of scripts. They wanted to check how high-scoring essays differ from low-scoring ones and came to the conclusion that there is not one dominant feature of high-scoring essays but a combination of them. This result was probably due to the fact that raters- as previously mentioned in section 2.4.2 on holistic ratings- seem to adopt a holistic rating style rather an analytic one. Their main findings were that the stronger predictors of IELTS scores were the word length of essays and low error rate. There seemed to be less errors (frequency of errors) in high-scoring scripts than low-scoring ones. This was also one of the findings in a small study by Demetriou 2004- see section below on vocabulary measures and teacher ratings. However, word count had a stronger correlation than any of the error categories, therefore it is one of the strongest predictors of band score in the IELTS Writing Task 2 performances. 'The average word count of high-scoring scripts was 336.9, compared to 265.8 for low-scoring scripts' (Mayor et al., 2002:256). Calculation of the TTR for each task showed that there was no apparent relationship between the different band levels and the TTR which is not a surprising result due to the fact that TTR is considered to be flawed (see Section 2.4.3). However, there is apparently a relation between the amount of speech (raw number of types, raw number of tokens) and the bands. Even though this result should be treated cautiously due to the small sample size, it was later supported by research from Banerjee et al. (2004).

In 2004, Banerjee, Franceschina and Smith investigated the different features of written language production at different IELTS band scores, using a large sample of 275 test participants. One of the main aspects researched was vocabulary richness. They suggested that counting the total number of words in a text (tokens) and the total number of different words in a text (types) is the simplest measure of lexical

richness. They also calculated the TTR (see below) but, due to the fact that it is considered to be a flawed measure (see section 2.4.3), it did not produce the expected results. Their research showed that there was a correlation between the number of tokens and types with the IELTS overall scores (IELTS band scores). They propose that the higher the band the candidate achieves in the IELTS exam, the higher the number of tokens and types in their speech. This makes sense because the more someone speaks they will produce more tokens and less types (as words will tend to be repeated). The researchers (Banerjee et al., 2004) also looked into the lexical sophistication of the test takers (the number of unusual words and the number of low or high frequency words used by the candidates). The Range programme was used to measure lexical sophistication. After Banerjee et al. (2004) calculated the candidates' lexical sophistication (as defined by the use of less-frequent words), it was established that the more advanced students used less high-frequency words and more infrequent words than less advanced students. In addition, they measured the lexical density of the candidates' written production. They defined lexical density as a measure which 'calculates the proportion of lexical words to grammatical words in the text' (O'Loughlin, 2001). The results showed that lexical density increased as the IELTS band levels increased. However, even though their research suggested that there are strong predictors for IELTS scores, the results should not be oversimplified and over generalised because there seems to be a multifaceted relationship between the variables that were investigated. Lastly, as previously mentioned and discussed in Section 2.6.2, Hawkey and Barker (2004) found that at higher IELTS proficiency levels essays were longer and employed with broader vocabulary**.**

2.9 VOCABULARY MEASURES AND TEACHER RATINGS

The relationship between vocabulary measures (and other aspects) and teacher ratings/scores has been investigated for years. A presentation of some of the major studies regarding this relationship will follow.

To begin with, in 1994 the Douglas study did not reveal a high correlation between test scores and the language produced by the learners. However, this claim was later rejected by other researchers such as Engber. Engber (1995) investigated the extent to which raters take lexical richness into account when rating learners' compositions. A high significant correlation was found between the scores and lexical variation (and

also for lexical variation minus error). Laufer et al. (2004) suggested that a single variable such as vocabulary size could be enough to predict academic scores. They state:

> 'on the other hand, vocabulary size on a single modality (such as 'passive recognition') may suffice as a surrogate measure of overall proficiency or as a predictor of academic performance, since a score on one modality, is likely to correlate highly with a score of any of the others' (Laufer et al., 2004:224).

The statement that vocabulary size has a correlation with teacher ratings/scores is reinforced by a study by Morris and Cobb who argue that 'the findings of the study reveal that the students' vocabulary profile results correlated significantly with grades'(Morris and Cobb, 2004:75). They used VocabProfile (Cobb), which is an online adaptation of Heatly, Nation and Coxhead's (2002) vocabulary assessment instrument. The correlations of the VocabProfile and grades were low to be used alone for assessing the learners, but could be used in combination with other aspects. Furthermore, vocabulary development, one of the six traits investigated in Lee et al.'s study (2009), was strongly correlated with the holistic scores. Essay length was also strongly correlated with the holistic score.

A major study by Daller and Phelan (2007) investigated the relationship between teacher ratings of EFL essays and the different aspects of lexical richness. Essays by 31 students studying EAP (English for Academic Purposes) were transcribed and then analysed using a mixture of measures of lexical richness such as TTR, D, Guiraud, P_Lex and Guiraud Advanced. The essays were rated by 4 EFL teachers using a set of IELTS band descriptors. The results showed that lexical sophistication in written essays influences teacher ratings more than lexical diversity (the use of advanced/rare words influences teacher ratings). The findings showed highly significant correlations between the teacher ratings and all the measures of lexical sophistication (measures that focus on non-frequent words). A possible interpretation of this result could be that teachers focus on advanced/rare words because they are easier to spot and count in the essays, thus saving them time. As a result, this could be the most 'economic' marking strategy for teachers. This result confirms the result of a previous study by Malvern and Richards (2002; Malvern et al 2004:103) that suggests that the use of advanced or rare words (lexical sophistication) influences the teacher ratings

of oral texts. Lorenzo-Dus's research also shows that lexical sophistication (the use of rare words) correlates with examiner ratings. She states that: 'a pattern could be identified whereby the candidates in the high- scoring bands produced more rare words within stretches of spontaneous talk than their low scoring band counterparts'(Lorenzo-Dus, 2007:228). Low-scoring candidates produced less rare vocabulary.

Demetriou's (unpublished Linguistics Project, 2004) research also showed that lexical sophistication (the use of infrequent words) was more important for EFL teachers than lexical diversity. However, the main finding was that errors, especially spelling errors, were more important to teacher ratings. Magnan's study (1988) examined the relationship between grammatical errors (different types) in oral proficiency interviews and oral proficiency ratings; it was found that there is a significant but not always linear relationship between them. Two of the main findings of this study shows that the relationship (between errors and ratings) is affected by the category of error, and also that learners tend to make more errors at higher levels as they become confident using more complex notions .

However, even though there are studies that suggest that lexical sophistication could have a higher correlation with teacher ratings than lexical diversity, this finding seems to be challenged in recent studies. Crossley et al. (2011a: 562) 'found that lexical diversity, word hypernymy values and content word frequency explain 44% of the variance of the human evaluations of lexical proficiency in the examined writing samples. The findings represent an important step in the development of a model of lexical proficiency that incorporates both vocabulary size and depth of lexical knowledge features'. For their study they used Coh-Metrix (a software tool) and TOEFL scores. According to Crossley et al. (2011b:190), 'Lexical diversity was the most predictive index and explained over 45% of the human ratings. Thus, the diversity of words in a sample best explains human judgements of lexical proficiency with high lexical proficiency samples contacting a greater variety of words'. Crossley et al. (2011a:574) state: 'Perhaps the most robust finding of this study is that an index of lexical diversity, D, explains almost 34% of the variance in human judgements of written lexical proficiency'. G. Yu also states that 'D had a statistically significant and positive correlation with the overall ratings of both writing and speaking

performances as well as the candidates' general language proficiency' (G Yu, 2009:236). G Yu's (2009) results revealed that D seemed to be a better predictor of speaking performance than writing performance. It could predict better speaking ratings than writing and males than females. G Yu states that: 'D was a significant predictor for the overall quality rating of compositions. However, other lexical features such as the number of types, tokens, short and long words, and average word and sentence length may also exert similar effects. In particular, the number of types and the number of long words seemed to be the other two most illuminative indicators, besides D, for the overall quality of the compositions. Together with D, they were able to predict a fairly large amount of the variances in overall quality rating' (2009:249).

Moreover, Ruegg, Fritz and Holland (2011) argue that lexical accuracy was predictive of lexis scores, but states that it is very hard to distinguish between lexis and grammar in ratings.

> 'In this test which was administered to incoming university students at the beginning of the academic year, it was found that lexical accuracy is predictive of lexis scores. The lexis scores, however, are predicted by the scores on the grammar scale much more than range, frequency, or even accuracy of lexis in the essays. The difficulties in separating lexis from grammar when rating writing are discussed.' (2011:63)

Based on previous research regarding the relationship between measures of lexical richness and teacher ratings, my investigation focuses on the prediction of teacher ratings based on measures of lexical richness. My research aims to confirm the results from previous studies and go a step further by investigating the relationship of various measures of lexical richness with teachers/examiners ratings (IELTS scores) by using 2 sets of data: a set of data taken from Greek-Cypriot EFL learners preparing for the IELTS examination (Study 1) and another set of data from Arab EFL learners (Study 2).

CHAPTER 3 – STUDY 1 / PILOT STUDY

This is the first study of the thesis which uses complete original data that I collected from Greek-Cypriot students preparing for the IELTS examination. Both oral and written data were collected in the form of speaking interviews and written essays in order to investigate the relation between measures of lexical richness and IELTS teacher ratings. This chapter introduces the research questions/hypotheses of Study 1 and presents and discusses results and limitations of the study.

3.1 RESEARCH QUESTIONS

1.   Which measures of lexical richness correlate highly with the teacher ratings? Will word-list based measures/lexical sophistication measures (such as Guiraud Advanced and P_Lex) correlate higher with the teachers' ratings than measures of lexical diversity (such as TTR and D)?
2.   To what extent can teacher judgement (this refers to global/holistic ratings) of IELTS essays and oral interviews be predicted by measuring the lexical richness of these texts?

3.2 METHODOLOGY

3.2.1 Participants

The subjects were 42 Greek-Cypriot students from 5 private schools in Cyprus that were learning English as a foreign language. All the participants were advanced level students preparing for the IELTS exam.
The sample was selected randomly. There were 22 male and 20 female students. Their age ranged from fifteen to eighteen years old. Their socio-economic status varied. None of them had ever lived in an English speaking country, all of them went to public schools, and none of them used English at home (Greek is their L1). They had all been learning English as a foreign language for approximately eight to nine years. All this information was collected by individual questionnaires given to each student before the start of the study.
They were all preparing for the IELTS examination for a period of almost one academic year (before the data collection).

3.2.2 Data Collection

The data collection took place in Cyprus in an English private school during the academic year 2008-2009. All the required forms were sent to the University's Ethics Committee and permission to go ahead with the research was granted. Data was collected at the end of the academic year (June 2009), two weeks before the participants took their IELTS Academic exam. They were all given a consent form to sign before the start of the study and an information sheet which explained the purpose of the study (see Appendix 1). All of the students were asked to complete a questionnaire which helped me obtain important information such as their age, sex etc.

The data collection was completed in two phases. On day one the participants were asked to write an essay under controlled exam conditions (mock exam), just like they would be asked to do in the real IELTS exam. The writing test consisted of two tasks. The complete format of the writing test is explained in Chapter 2 under the IELTS Test section. I only chose the second task to include in my study to measure the candidates' vocabulary size because the first task would not be ideal for this. From my own experience of teaching IELTS, certain expressions are commonly learned by students and most students use the same words and expressions when describing graphs, diagrams etc.

The candidates were only asked to do Writing Task 2, which is the main part of the writing exam (not Task 1). According to research, Task 2 in writing 'places greater textual demands on candidates than Task 1' (Mickan and Slater, 2003:61). In addition, 'IELTS examiners give more weight to Task 2 in marking than Task 1' (Uysal, 2010:315).

The topic was selected from an IELTS past exam paper (Official IELTS Practice Materials 2003). The card for the essay is displayed below:

Writing task

You should spend about 40 minutes on this task.

Present a written argument or case to an educated non-specialist audience on the following topic:

Modern technology is transforming the way we work and is of benefit to all of society.

You should write at least 250 words.

You should use your own ideas, knowledge and experience to support your arguments with examples and relevant evidence.

On day two the participants were asked to return to their schools and give an oral interview lasting fifteen minutes, exactly as they would in the real IELTS exam (Speaking Test). The terms *speaking test* and *interview* are interchangeable here because the IELTS speaking test is conducted in the form of an interview. The person conducting the interviews was a trained IELTS examiner; this helped ensure the conditions were as realistic as possible. The raters/examiners were trained accordingly. The IELTS Speaking exam consists of three stages: In Stage 1 the examiner asks the candidate personal questions (the candidates introduce themselves; tell the examiner about their families and their hobbies etc.). In Stage 2 they are given a topic card and they have to speak for 2 minutes without interruption. In Stage 3 follows an interaction with the examiner and the candidate. In this part the candidate is asked questions related to the topic on the card (from Stage 2), but the questions are more abstract (see Section 2.5 –Chapter 2 for a full description of the IELTS Speaking Test). The topic used for Stage 2 of the exam was selected from a past exam paper (Official IELTS Practice Materials, 2003). This is shown below:

Picture 3.2. IELTS Speaking Part 2 Topic Card

> Describe a wedding you have been to or heard about.
>
> You should say:
>
> - who got married
> - what they wore
> - what they did on the day
>
> and explain how you felt about this wedding.

3.3 MEASURES AND PROCEDURES

3.3.1 Transcriptions

The essays and oral interviews of all 42 students were then transcribed into CHAT (MacWhinney, 2000) to help compute the measures of lexical richness. CHAT prepares the texts (with the addition of various symbols and coding) for the analysis. Spelling was corrected in order to avoid misspelled words that would not be recognised by the programme being counted as advanced words. Some words, such as place names, were excluded from the calculations to prevent them from being counted as advanced/sophisticated words. Words that were double (words that were spoken or written twice in error) were also excluded from the calculations for vocabulary size. All the essays and interviews were typed before being given to the raters. Below is an example of one of the transcriptions:

@Begin

@Languages: en

@Participants:	KYR Anonymous student
*KYR:	Surely modern technology take an important role in the way we work and in what way, that, help us in our jobs.
*KYR:	With the transform of the technology the society have its own benefits.
*KYR:	There are a lot of benefits in the society and in our job by using Modern Technology.

*KYR:First of all, the technology make our jobs easier and sometimes help the people to have a more relaxing day in their works.
*KYR:About the economy of our country, by using modern technology is increasing.
*KYR:Furthermore modern technology make us find job easier.
*KYR:As long as the benefits of the modern technology in the society and in our countries are a lot.
*KYR:People can get more money with doing easier job.
*KYR:More people visiting Cyprus and the tourism is increasing.
*KYR:Finally unfortunately with modern technology our society and the way we are working have some disadvantages.
*KYR:The Cypriots sometimes loses their jobs from the tourist they are coming in Cyprus.
*KYR:In addition older people can not understand the development of the technology and they find that it is not a special thing but it is a new event in our daily lives.
*KYR:To sum up for the young people and especially for as the teenagers modern technology is one of the most important thing in our lives because it is making our future.
*End

The essays were rated by two trained IELTS Examiners in the UK. It needs to be clarified that the terms teachers/raters/examiners are used interchangeably in this study because all the examiners/raters used were also teachers. However, it needs to be acknowledged that these terms are not necessarily the same thing (not all teachers are IELTS or any other exam trained). The oral interviews were also rated on the spot by the examiner who conducted the interviews. Instructions were given to all raters explaining exactly what was asked from their part. For the overall mark of the essays, the teachers were asked to use the IELTS Overall Band Score.

The raters also had to mark the interviews (oral data), giving an overall mark by using the IELTS Overall Band Descriptor. The examiners had to rate the essays and interviews on a 9-point scale, with 9 being the highest mark that showed greatest language proficiency.

A description of the IELTS band descriptors used is provided on the following pages:

Picture 3.3. IELTS Overall Band Descriptor

IELTS Band Descriptors	
Band	**Descriptor**
9	Expert user Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.
8	Very good user Has fully operational command of the language with only occasional unsystematic inaccuracies. Misunderstandings occur in unfamiliar situations. Handles complex detailed argumentation as well.
7	Good user Has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.
6	Competent user Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.
5	Modest user Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.
4	Limited user Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language.
3	Extremely limited user Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur.
2	Intermittent user No real communication is possible except for the most basic information using isolated words or short formulae in familiar

	situations and to meet immediate needs. Has great difficulty in understanding spoken and written English.
1	Non user Essentially has no ability to use the language beyond possibly a few isolated words.
0	Did not attempt the test No assessable information provided.

(IELTS Band Descriptors, http://www.ielts.org)

The raters/examiners were also interviewed by me to explain any unusual marks and make any further comments. They also commented on each essay and interview if they felt it was necessary. In the actual exam examiners are asked to comment on the number of words, whether the essay was under length or off-topic and if it was memorised or illegible. The interviews were conducted to gain an insight of what the examiners were thinking during rating the data and check how they decided to award specific band scores. These interviews were not used for analysis. To add a hint of qualitative analysis (a qualitative aspect) to my study I also requested that the raters made a note if there was something in particular that influenced their decision for a specific mark. After the quantitative analysis of the data (essays and students interviews/speaking tests) I looked at the raters/examiners' notes/comments (that were written on each essay or speaking test transcription) to check for any patterns or justification for some of their decisions.

1 point was deducted for under length essays. For example, if the band allocated was 6, the rater would make it a 5 because of the essay being under the amount of words they were asked to write. One could argue that this instruction contrasts with the general view that quality in academic writing can result from careful use of words and grammar to produce more precise and concise sentences. However, this is something that IELTS trainers are instructed to do in the exam. Candidates are penalised for shorter word counts.

3.3.2 To Lemmatise or not to lemmatise?

In the previous chapter a major issue was presented regarding the problematic nature of what constitutes a word and what should be counted as a word. Several questions then arise when conducting research: Should we lemmatise data or not? Should we count word families? It depends on what we choose to count as words. The results can be affected by this decision (Knowles and Don, 2004). There are researchers who suggest that data should be lemmatised. According to Coxhead, learners do not make much effort to understand an inflected or derived member of a family if they are familiar with the base word (Coxhead, 2000). Therefore, in Coxhead's study words were defined as word families. Beglar (2010) also argues that the word family can be used as a vocabulary measure due to the fact that more proficient learners should be able to identify words and use word building devices. Treffers-Daller (2013) also highlights the importance of lemmatising the data because it can increase the explanatory power of lexical richness measures, especially for highly inflected languages such as French.

On the other hand, the following researchers argue for the importance of non-lemmatisation of data. Knowles and Don (2004: 71) state that 'generalizations about whole lemma become less and less convincing' as detailed linguistic examinations of corpus-based data continue to be performed, and that researchers may need to begin 'to consider individual words' or 'actually even individual word meanings' as the basis for their analyses. In G Yu's (2009) study, the inflections of the same word were treated as different types for the reason that lexical diversity was analysed as an end-product, and in IELTS ratings the candidates need to demonstrate 'accurate morphological word forms control'. In addition, Schmitt and Zimmerman's study (2002) shows that learners have difficulty understanding all the derivative forms of a word (especially adjectives and adverbs) therefore we should not assume that because a learner knows a word that they should be familiar with all the different derivative forms of that word. 'The results indicate that knowledge of one word in a family does not necessarily imply productive knowledge of other forms in that family' (2002:162). According to Beglar and Hunt (1999:149), 'knowledge of a word's base form does not guarantee knowledge of its derivatives or inflections'. Therefore, the decision for non-lemmatisation of data in the present study is justified by the

researcher's wish to check for accurate word formation, which is one of the aspects mentioned in the IELTS band descriptors, and an important process during language learning. In addition, according to Broeder and Voionmaa (1985) lemmatisation is time-consuming and does not give you any additional information.

3.3.3 The Lexical Richness selected measures

For the measurement of lexical richness the following measures were included:

Lexical Sophistication Measures	Lexical Diversity Measures	Raters' Judgements
Number of types	Number of tokens	IELTS Written Overall Band Score
Guiraud Advanced	Guiraud	IELTS Oral Overall Band Score
P_Lex (Lambda values)	TTR (Type-Token Ratio)	
	Malvern and Richards D	

The measures under the first two columns (lexical sophistication and lexical diversity) are objective measures as they are based on mathematical models or are computer based whereas the raters judgements (IELTS band scores) are subjective measures.

A description and justification of the selection of the measures will be provided here but for a more detailed discussion on each measure please refer back to Chapter 2 (Sections 2.4.3 and 2.4.5).

Number of tokens

This is the total number of words in the essays or interviews (speaking tests). This was included as it is considered one of the simplest measures of vocabulary size and previous research (Banerjee et al., 2004) revealed that it correlates highly with teacher/examiner ratings.

Number of types

This is the number of different words used in an essay. It was included due to indications from previous studies (Banerjee et al., 2004) that it can act as a predictor of teacher ratings.

TTR

Type-Token Ratio. The TTR was included (despite its flaws) as it is an old and established measure. It was not included in Turlik's (2008) study or Read and Nation's (2002) and this was something I wanted to further investigate. I wanted to check if TTR would be discarded from my predictive model or if it would help my model improve.

Guiraud

A mathematical transformation of the TTR in order to improve the text length problem. Guiraud is calculated by dividing Type by the square root of tokens. Guiraud was included because it was not included in Read and Nation's (2002) study either.

D

D is calculated by the *vocd* command in CLAN. This measure was also designed to overcome the text length effect and it was included because it overcomes problems with text length. This was included in order to have comparable results with Read and Nation.

Guiraud Advanced

Guiraud Advanced was also selected as a measure of lexical sophistication because, according to research (Daller et al, 2007), it is a valid measure.

P_Lex

P_Lex was chosen over Lex30 because Lex 30 is a single item test whereas LFP and P_Lex are measures of texts and more suitable for my study. It was also chosen over LFP for reasons discussed in Chapter 2 (Section 2.4.5).

3.3.4 Equipment and Software

The *vocd* command was used in CLAN to calculate the number of types, tokens, TTR and D values of the essays. Words that were repeated or place names were excluded from the count. All ratings were put in an SPSS file along with the scores of lexical

richness of both lexical diversity and sophistication measures. The SPSS file consisted of the following variables:

Names of students

Gender

Written and Oral Overall Marks by each IELTS examiner

'D' value

Number of types

Number of tokens

Guiraud

Guiraud Advanced

TTR

P_Lex

For the calculation of all the lexical measures for the oral data, only the second stage of the speaking test was used, as herein the examiner speaks uninterrupted.

Guiraud Advanced was calculated by using Eugene Mollet's programme (personal communication with Daller) and was based on two wordlists. The base wordlists that were used in this study for the measurement of lexical richness (in order to calculate Guiraud Advanced) are:

Base list 1: This list is based on the first thousand words (ranked according to frequency) of West (1953).

Base list 2: Based on the next thousand words of the Paul Nation's word list (See Nation, URL), which is based on West (1953).

The purpose of the word lists is to identify rare words for the lexical sophistication measures. Each word that cannot be found in Base List 1 or 2 will be counted as an advanced/rare word. Both lists were ticked for the calculations. Each text/essay was uploaded and a value was given automatically. Here follows a short description of the above word lists:

West's Service List of English words GSL (1953) consists of the two thousand most useful word families in English. The words represent the most frequently used words in English and were selected from a corpus of written English (Daller and Phelan, 2007).

Even though the GSL list has been criticised for many reasons, research into academic texts by Coxhead (2000) has shown that it is reliable because it covers almost 80% of the words of the academic texts she studied. Therefore, it is essential for any EAP student to know these word families (Gillet, www.uefap.co.uk).

P_Lex was used to calculate lambda values. P_Lex is a computer programme that models the occurrence of rare words in a text. The dictionary needed for the programme to work is based on Paul Nation's word lists (Xue and Nation, 1984). Lambda values normally range from 0 to 4.5, and the higher the figure, the higher the proportion of infrequent words. Each text was checked before the report and words were checked individually. All the words that were in Level 0 and Level 1 word lists were considered easy words, whereas words that did not belong in these two lists (and were not proper names, mistakes or numbers) were considered sophisticated words.

3.4 DATA ANALYSIS AND RESULTS

3.4.1 Descriptive statistics

For this study there were 42 participants: 20 female students and 22 male students. The descriptive statistics for measurements related to written data are presented. It can be seen from Table 3.1 that the mean number of Types is 119.74, the minimum is 78, and the maximum is 149. Regarding the variables names in the tables below, the letters *Wr* is an abbreviation of the word *written* and were added next to each variable that refers to written data. The letters *Or (oral)* were added in all variables that refer to oral data.

Table 3.1: Descriptive statistics for measurements related to written data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
D Written Data	42	44	121	81.70	19.243
Types Wr	42	78	149	119.74	16.510
Tokens Wr	42	119	318	226.17	38.776
TTR Wr	42	0	1	.54	.066
Guiraud Wr	42	6	9	7.99	.793
Guiraud Adv Wr	42	0	2	.89	.362
P_Lex Wr	42	1	3	1.47	.375
Valid N (listwise)	42				

Below, the descriptive statistics for measurements related to oral data are presented. It can be seen from Table 3.2 that the mean number of Types is 77.21, the minimum is 38, and the maximum is 111.

Table 3.2: Descriptive statistics for measurements related to oral data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
D Oral data	42	35	122	54.36	13.975
Types Or	42	38	111	77.21	19.962
Tokens Or	42	58	278	146.88	56.484
TTR Or	42	0	1	.55	.085
Guiraud Or	42	5	8	6.41	.598
Guiraud Adv Or	42	0	1	.89	.247
P_Lex Or	42	0	1	.48	.294
Valid N (listwise)	42				

3.4.2 Inferential statistics- hypothesis testing

Cronbach's Alpha coefficient is one of the most useful tools for checking the reliability of a scale and generally the inter-rater reliability (Field, 2005). In general, Cronbach's Alpha coefficient scale needs to be over 0.7 in order to be reliable with the sample (Nunnally, Durham, Lemond and Wilson, 1975). Therefore, reliability (Cronbach's Alpha) for the written and oral scores has been calculated. Table 3.3 presents the number of items, the mean, the standard deviation, and the final Cronbach's Alpha coefficient for each factor. It can be seen from Table 3.3 that

Cronbach's Alpha value is less than 0.70 for the written overall case. These values show the high extent to which a scale produces consistent results if repeated measurements are made on the characteristics. This is an important limitation for this study which may have been caused by the small sample size or realistic differences between the examiners. Further results are presented in Appendices 3 and 4 for the written and oral scores, respectively.

Table 3.3: Reliability Statistics

Reliability Statistics					
	N of Items	Mean	Std. Deviation	Cronbach's Alpha	Cronbach's Alpha Based on Standardised Items
Written Overall	2	10.833	0.973	0.578	0.584
Oral Overall	3	16.548	1.692	0.795	0.800

What follows is the paired samples t-test for two dependent samples, performed to test for any significant differences among the scores of the two examiners used for this study, for the written scores. First of all, the assumption that the paired differences should be normally distributed is tested in Appendix 5 for the differences of the written scores, between the first and the second examiner. The Kolmogorov-Smirnov and the Shapiro-Wilk tests were used to test normality. In both cases, these tests suggest that normality cannot be assumed for all factors. However, using the Central Limit theorem, as the sample size is large enough (more than 30), the mean of each factor can be assumed to be approximate to the normal distribution. So, both parametric (paired samples t-test) and non-parametric tests (Wilcoxon Signed Rank test) will be used for data analysis. It can be seen from Table 3.4, that there are statistically significant differences for the written overall rating. In other words, responses seem to be scored statistically higher by the second examiner. In addition, it can be seen that both the parametric and non-parametric tests, suggest the same conclusions.

Table 3.4: Means, standard deviations and t-values derived from comparisons between the examiners' overall ratings for written data

	Mean	S.D.	t	df	P-value	Wilcoxon P-value
Wr Overall	-0.45	0.63	4.63	41	< .001	< .001

Kendall's W (coefficient of concordance) is a measure of the agreement of the rankings of variables across cases. Below is Kendall's W test used to test for significant differences among the scores of the three examiners that were used for the oral scores in this study. The one way analysis of variance was not used, as it is not proper for related samples. It can be seen from Table 3.5, that Examiner 1 and 3 have lower ranks, which indicate lower scores compared to Examiner 2. The Kendall's W is equal to 0.5 which indicates moderate agreement in the ordering across cases. The highly significant value of p (<.001) indicates that at least one of the examiners scores differs from the others. It can be concluded that Examiner 2, differs from the others.

Table 3.5: Kendall's W Ranks and Test

	Mean Rank Or Overall
EX1	1.62
EX2	2.79
EX3 (main)	1.6
N	42
Kendall's W	0.50
Chi-Square	42.17
df	2
Asymp. Sig.	0.00

The mean written overall scores for the two examiners and the mean oral overall scores were used for further analysis. Only the overall holistic score for the model was used because, according to the literature (Malvern, et al., 2004;), when holistic rating is used raters give the same rating as they give to most of the separate traits in a scale. This is called the 'halo effect' (see more in Glossary and Terms, Section 2.7.5). In addition to this, instead of the actual values of the 'D Written Data', the 'Types Wr', the 'Tokens Wr', the 'D Oral Data', the 'Types Or' and the 'Tokens Or', their natural logarithm was used for further analysis via correlation and regression (which is now represented by the letters *Ln* in front of the variables- see tables below). The natural logarithm was used to transform the data, in order to create new values that are nearer to written and oral mean overall scores (that theoretically take values from one up to nine). Also, in this way, the assumption of linearity among depended and independent variables and the assumption for normality of residuals in linear regression, has been improved. Lee et al. (2009:389) suggest that sometimes data needs to be transformed

'…since some of these ratio variables often turn out to have extremely small variances, these variables are usually mathematically converted to more statistically stable values (by way of logarithmic transformation…)'.

In Table 3.6 below you can see the Spearman's Rho correlation coefficients for the seven independent measurements of this study, and the written overall mean score of the two examiners. Table 3.6 suggests a positive significant relationship between the written score and the natural logarithm of the written types (r_{42}= 0.335, p<0.05). Also, there is a strong positive significant relationship between the written score and the natural logarithm of the written Guiraud Adv Wr (r_{42}= 0.322, p<0.05) and the P_Lex Wr (r_{42}= 0.328, p<0.05). These findings were expected, and are supported by the literature (Daller et al., 2007). Furthermore, these findings enable us to use multiple linear regression to predict the overall mean written score.

Table 3.6: Spearman's Rho Correlation Matrix

Correlations

		Wr Overall	Ln(D Written Data)	Ln(Types Wr)	Ln(Tokens Wr)	TTR Wr	Guiraud Wr	Guiraud Adv Wr	P_Lex Wr
Wr Overall	Correlation Coefficient	1.000	.189	.335*	.174	-.004	.288	.322*	.328*
	Sig. (2-tailed)	.	.231	.030	.269	.978	.065	.037	.034
	N	42	42	42	42	42	42	42	42
Ln(D Written Data)	Correlation Coefficient	.189	1.000	.570**	-.155	.776**	.844**	.280	.153
	Sig. (2-tailed)	.231	.	.000	.329	.000	.000	.073	.334
	N	42	42	42	42	42	42	42	42
Ln(Types Wr)	Correlation Coefficient	.335*	.570**	1.000	.622**	.144	.853**	.065	.067
	Sig. (2-tailed)	.030	.000	.	.000	.364	.000	.684	.673
	N	42	42	42	42	42	42	42	42
Ln(Tokens Wr)	Correlation Coefficient	.174	-.155	.622**	1.000	-.617**	.150	-.250	-.126
	Sig. (2-tailed)	.269	.329	.000	.	.000	.344	.110	.425
	N	42	42	42	42	42	42	42	42
TTR Wr	Correlation Coefficient	-.004	.776**	.144	-.617**	1.000	.606**	.390*	.337*
	Sig. (2-tailed)	.978	.000	.364	.000	.	.000	.011	.029
	N	42	42	42	42	42	42	42	42
Guiraud Wr	Correlation Coefficient	.288	.844**	.853**	.150	.606**	1.000	.257	.204
	Sig. (2-tailed)	.065	.000	.000	.344	.000	.	.100	.196
	N	42	42	42	42	42	42	42	42
Guiraud Adv Wr	Correlation Coefficient	.322*	.280	.065	-.250	.390*	.257	1.000	.687**
	Sig. (2-tailed)	.037	.073	.684	.110	.011	.100	.	.000
	N	42	42	42	42	42	42	42	42
P_Lex Wr	Correlation Coefficient	.328*	.153	.067	-.126	.337*	.204	.687**	1.000
	Sig. (2-tailed)	.034	.334	.673	.425	.029	.196	.000	.
	N	42	42	42	42	42	42	42	42

*. Correlation is significant at the 0.05 level (2-tailed).

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

As displayed below, the multiple linear regressions were used to test if the independent variables (measurements) were related and could explain the overall mean written score. The stepwise method has been used, and the best selected model using this method is presented below. Furthermore, the full output of the regression analysis is presented in Appendix 6.

Table 3.7: Model Summary of Regression Analysis for Written Overall Score

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
					R Square Change	F Change	df1	df2	Sig. F Change
2	.474	.224	.185	.43939	.102	5.138	1	39	.029

The two independent variables [P_Lex Wr and Ln (Tokens Wr)] can explain 22.4% of the written overall score (R^2=0.224). It can be seen from the ANOVA Table 3.8 that this model is significant ($p< 0.01$) which indicates that at least one of the independent variables (the lexical richness measures) helps explain the overall written score. The results indicate that the independent variables have unequal strength in explaining the written overall score. In addition to this, as shown in Table 3.9, the two independent variables are significant and positively related with the written overall score.

Table 3.8: ANOVA for the multiple linear regression model

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	2.179	2	1.089	5.643	.007(b)
2	Residual	7.529	39	.193		
	Total	9.708	41			

Table 3.9: Regression Coefficients

Model		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	Sig.
2	(Constant)	-.016	2.120		-.007	.994
	P_Lex Wr	.541	.187	.417	2.892	.006
	Ln(Tokens Wr)	.857	.378	.327	2.267	.029

It can be concluded that P_Lex Wr is significantly positively related with the overall written score (b=0.541, t=2.892, p-value<0.01). This variable has the greatest strength in explaining the overall written score (beta=0.417). The natural logarithm of the written tokens has the second greatest strength in explaining the overall written score (beta=0.327) and is significantly positive related with the overall written score (b=0.857, t=2.267, p-value<0.05). The fitted regression model (Table 3.9) is:

Overall written score= -0. 016 +0.541* P_Lex Wr+0.857*Ln (Tokens Wr).

In Table 3.10 below, the Spearman's rho correlation coefficients for the seven independent measurements of this study (related to the oral data) and the oral overall mean score of the three examiners are presented. It can be suggested that there is a positive significant relationship between the oral overall score with the natural logarithm of the oral types (r_{42}= 0.590, p<0.01), with natural logarithm of the oral tokens (r_{42}= 0.541, p<0.01), with Guiraud Or (r_{42}= 0.604, p<0.01) and P_Lex Or (r_{42}= 0.322, p<0.05). On the other hand, there is a strong negative significant relationship between the oral overall score with TTR Or (r_{42}= -0.430, p<0.05). These findings enable us to use multiple linear regression to predict the overall mean oral score.

Table 3.10: Spearman's Rho Correlation Matrix

Correlations

		Or Overall	Ln(D Oral data)	Ln(Types Or)	Ln(Tokens Or)	TTR Or	Guiraud Or	Guiraud Adv Or	P_Lex Or
Or Overall	Correlation Coefficient	1.000	.177	.590**	.541**	-.430**	.604**	.073	.322*
	Sig. (2-tailed)	.	.263	.000	.000	.004	.000	.647	.038
	N	42	42	42	42	42	42	42	42
Ln(D Oral data)	Correlation Coefficient	.177	1.000	.113	-.061	.344*	.540**	.161	.044
	Sig. (2-tailed)	.263	.	.476	.701	.026	.000	.307	.782
	N	42	42	42	42	42	42	42	42
Ln(Types Or)	Correlation Coefficient	.590**	.113	1.000	.973**	-.805**	.803**	-.078	.307*
	Sig. (2-tailed)	.000	.476	.	.000	.000	.000	.623	.048
	N	42	42	42	42	42	42	42	42
Ln(Tokens Or)	Correlation Coefficient	.541**	-.061	.973**	1.000	-.909**	.667**	-.115	.262
	Sig. (2-tailed)	.000	.701	.000	.	.000	.000	.468	.094
	N	42	42	42	42	42	42	42	42
TTR Or	Correlation Coefficient	-.430**	.344*	-.805**	-.909**	1.000	-.372*	.195	-.166
	Sig. (2-tailed)	.004	.026	.000	.000	.	.015	.216	.293
	N	42	42	42	42	42	42	42	42
Guiraud Or	Correlation Coefficient	.604**	.540**	.803**	.667**	-.372*	1.000	.093	.375*
	Sig. (2-tailed)	.000	.000	.000	.000	.015	.	.559	.015
	N	42	42	42	42	42	42	42	42
Guiraud Adv Or	Correlation Coefficient	.073	.161	-.078	-.115	.195	.093	1.000	.321*
	Sig. (2-tailed)	.647	.307	.623	.468	.216	.559	.	.038
	N	42	42	42	42	42	42	42	42
P_Lex Or	Correlation Coefficient	.322*	.044	.307*	.262	-.166	.375*	.321*	1.000
	Sig. (2-tailed)	.038	.782	.048	.094	.293	.015	.038	.
	N	42	42	42	42	42	42	42	42

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

*. Correlation is significant at the 0.05 level (2-tailed).

As shown below, the multiple linear regressions were used to test if the independent variables (measurements) were related and could explain the overall mean written score. The stepwise method has been used, and the best selected model using this method is presented below. Additionally, the full output of the regression analysis is presented in Appendix 7.

Table 3.11: Model Summary of Regression Analysis for Oral Overall Score

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.607(a)	.368	.353	.45389	.368	23.325	1	40	.000

The independent variable selected [Guiraud Or] can explain 36.8% of the oral overall score (R^2=0.368). It can be seen from the ANOVA Table 3.12 that this model is significant (p< 0.01).

Table 3.12: ANOVA for the multiple linear regression model

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.805	1	4.805	23.325	.000(a)
	Residual	8.241	40	.206		
	Total	13.046	41			

It can be concluded that Guiraud Or is significantly positive related with the overall oral score (b=0.572, t=4.830, p<0.01). The fitted regression model is:

Overall oral score= -1.845 +0.572* Guiraud Or.

Table 3.13: Regression Coefficients

Model		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.845	.763		2.417	.020
	Guiraud Or	.572	.118	.607	4.830	.000

3.5 DISCUSSION

In this section each of the hypotheses/research questions that were introduced at the beginning of Study 1 are addressed and discussed. The first question was the following:

1.   Which measures of lexical richness will correlate highly with teacher ratings? Will measures of lexical sophistication correlate higher than measures of lexical diversity?

I expected to find that measures of lexical sophistication (measures based on word-lists) would correlate higher with the ratings than measures of lexical diversity. This was confirmed by the written data because it seems that, from the analysis of the data, the variables that had higher correlations with the written overall score were the types, Guiraud Advanced and P_Lex. Guiraud Advanced and P_Lex are both measures of lexical sophistication so I expected them to correlate highly with the examiner ratings for the essays. These results are also supported by the literature (Banerjee et al., 2004; Daller and Phelan, 2007). The results regarding the oral data were not as expected. It was found that there is a strong positive significant relationship between the oral overall score (given by the examiners) and the types, tokens, Guiraud and P_Lex. Subsequently, it can be seen that not only were measures of lexical sophistication (P_Lex) highly correlated with the scores, but also with measures of lexical diversity. There is also a strong negative significant relationship between the oral overall score and the TTR (which is another measure of lexical diversity and, according to many researchers in the literature, quite a flawed one).

A further questions was:

2.   <u>To what extent can teacher judgement (this refers to global/holistic ratings) of IELTS essays and oral interviews be predicted by measuring the lexical richness of these texts?</u>

From the regression it was found that the two independent variables that can explain 22.4% of the written overall score are the tokens and P_Lex (one variable from lexical diversity and one lexical sophistication). As for the model for predicting the oral overall score, it seems that Guiraud is the only independent variable that can explain 36.8% of the score, which is not what was expected and is not supported by the literature. From what is suggested in the literature, measures of lexical sophistication such as Guiraud Advanced or P_Lex should be better predictors of scores for oral data. One possible explanation for the different results (different variables) regarding the oral and written data could be the nature of the tasks. Written tasks usually require the use of more formal language (therefore, more sophisticated/infrequent words). Thus, P_Lex, which is a measure of lexical sophistication, was found to be one of the predictors of the written ratings, whereas in oral data (where the use of language could be more colloquial) Guiraud, which is a measure of lexical diversity, was a better predictor of the ratings (see discussion on oral and written registers in Chapter 2).

Even though my hypothesis was partly confirmed by the written data, there were some aspects that could be improved in order to repeat the study and obtain better results. First of all, the reliability of my raters was not high (after being calculated using the Cronbach's Alpha). The low reliability of the raters proved a hindrance to the study. The inter-rater reliability could be massively improved if more raters/examiners were used to score the written and oral data. In the future, if a larger study is to be repeated and replicated a larger amount of examiners needs to be used. In addition, even though the participant/student sample was not small, using an even larger number of students would make the findings more reliable. However, finding 42 students and 3 examiners to participate in the study was difficult enough. Therefore, if the study was to be repeated, it would be ideal if I would be given permission to obtain data from the IELTS organisation and their massive IELTS database. A study could then be repeated with

larger amounts of participants, and other variables could be added to the model (from the information the organisation has for each student).

It was mentioned above (in the methodology section) that apart from asking the examiners to rate the data, I also asked them to write comments at the bottom of each test justifying the mark given (especially if it was an extreme score, i.e. something very low or very high). After thoroughly investigating the examiners' comments justifying the given marks/ratings for each essay, it was noticeable that the aspects that are found to be most 'off-putting' are grammatical errors. It would therefore be a good idea to repeat the study, count the number of grammatical errors, and add them as a variable to the model to check if it would make an improvement. There is an approach by Engber (1995) which measures the lexical errors in a text (percentage of lexical errors in a text). This approach could be used in further research as a means to improve my model for predicting IELTS band levels (IELTS scores). However, it should be noted that it is hard to define, identify and make a distinction of errors in analyses (Lennon, 1991). Furthermore, the model for predicting the written overall score could be further improved by using both parts of the Academic Writing Test. For this study I have only used Writing Task 2 because it contributes more to the total writing score and uses a larger variety of vocabulary (due to the nature of the task). Maybe it would be better to use both parts of the Writing Test to make sure that the examiners realise what the candidates' abilities are and test whether the model would be improved if the other parts of the exam were added.

In regards to the oral data, as was explained in the methodology section, all the vocabulary values obtained from using the different measures of lexical richness were calculated after transcribing only one section of the Academic Speaking Test. The reason behind this decision was the fact that Section 2 of the speaking exam is a monologue by the candidate (the candidate speaks for about 2 minutes without interruption). Therefore, the values (Lambda, D etc.) only represented that part, whereas the marks/scores given by the examiners were given after listening to the whole exam (15 minutes in total). This could be another reason that my model could only predict a certain percentage of the score. It could be improved if the whole exams are transcribed (15 minutes instead of 2), and the calculations are made based on these larger transcriptions.

What needs to be highlighted in the pilot study is the use of both oral and written data for the analysis. As already discussed in Chapter 2 (Section 2.3.5) there is a difference between using written and oral registers. Ratings of spoken fluency may reflect other traits than ratings of essays. For example, one could speculate that different accents (or a heavy accent) can influence raters' judgement when rating oral data. In addition, some aspects of lexical richness e.g. sophistication are more salient in oral speech. However I did not investigate this as my focus was different. I agree that 22.4% does not seem a very satisfactory score and further research would definitely give more insight to what the remaining percentage of the variance of the scores explains. If only 22.4% for the written overall and 36.8% of the oral overall of the variance in the ratings can be attributed to lexical knowledge then what remains of the percentage of the overall score may be explained by other variables or even social factors. It could maybe be explained by other variables such as under-performance of students or unfamiliar topics (see discussion on influence of topic and task in language testing- Section 2.5.6).

In addition, researchers need to be careful as to what measures to use for analysing written or oral data due to the fact that some measures seem to work best with written and some with oral data. Furthermore, the low percentage (22.4%) of the variance of the written ratings could be explained by my use of a free productive task which produces much variation in the data (not so controlled). Regarding rating oral data, Brown (2003) suggests that it can be influenced by other non-linguistic factors such as interviewer behaviour (such as compensating for less-than-competent interviewers).

Referring to percentages when describing teacher ratings may seem quite odd but it is quite useful for statistical purposes. What those percentages show is the ranking of importance of specific features. For example, when the results produced in the model represents vocabulary as 22% of the written ratings, this shows vocabulary is quite important in teacher ratings but not as important as for oral ratings in which 36.8 % of the variance in the ratings can be explained by that. Therefore, these numbers are useful regarding the creations of statistical models that are then open to interpretation by each researcher.

It is almost certain that a combination of more qualitative and quantitative analysis could improve the model. This is something that should be looked into in future research.

After the results of Study 1, which only used measures of vocabulary breadth as predictors of teacher ratings, it is acknowledged that the research needs to be taken a step further and should look at adding other measures of vocabulary knowledge (measures of depth of vocabulary) into the IELTS model to improve its predictive validity. Depth of vocabulary is a construct hard to define and operationalise but research (Beglar and Hunt, 1999; Qian and Schedl, 2004) suggests that the use of formulaic language (such as collocations or phrasal expressions) is an aspect of depth of knowledge. Therefore, it was decided to add the extra variable 'formulaic count' in the model. What follows in the next chapter is an introduction to what formulaic sequences are, how they can be operationalized, and why they are used in this study.

CHAPTER 4 –ADDING FORMULAIC SEQUENCES TO THE MODEL

4.1 INTRODUCTION

This chapter continues with the issues raised in Chapter 2 and aims to provide a more detailed discussion of MWU (multiword units) and formulaic language in particular. The chapter comprises 8 sections which consist of various subsections. I start from providing a definition of formulaic sequences and a discussion of their acquisition, use, teaching and learning and how they can be detected in a text. I then turn to different types of formulaic sequences such as idioms and phrasal verbs. Even though collocations are considered to be an example of formulaic sequences they are discussed in a separate section, since they are the main focus of Study 2. In this section I discuss definitions of collocations, along with their acquisition and use, and acknowledge the importance of the relationship between collocations and the frequency factor. Then follows a discussion on word lists and academic corpora. I next discuss the link between formulaic language (collocations in particular) and different aspects of L2 proficiency, including discussions of previous findings, the relationship between written and oral data, and teacher ratings. The next section presents various methodological problems that can be encountered when conducting research into formulaic language, specifically problems with data collection, before focusing on issues surrounding definitions and the operationalisation of terms and clarifying the way in which the term is used in this thesis. I conclude the chapter with the rationale of this study and operationalisation of the formulaic sequences term.

As already mentioned in Chapter 2, formulaic language plays a major role in language learning, teaching and testing. Schmitt (2010:9) argues that formulaic language holds a prominent place in vocabulary research. Kovesces and Szabo (1996: 328) state that 'the vocabulary of a language cannot be equated with the sum of the single words in the language'. Even though most people think of words when they hear the word *vocabulary* (Hill, 2000), the authors state that many corpus studies have shown that a large percentage of a text consists of multiword expressions or collocations (Nattinger and DeCarrico, 1992; Pawley and Syder, 1983).

According to Hyland (2008), research into formulaic patterns and sequences has been happening since 1924 (Jespersen), and then in 1952 Firth (titled 'the Father of British Linguistics' by Möbarg, 1997:204) made the term *collocation* popular. In recent times the importance of formulaic patterns (or lexical chunks) was highlighted by Nattinger and DeCarrico (1992), whose work Boers and Lindstromberg (2009) considered to be 'the milestone in the growth of appreciation of the place of chunks in language learning' (2009:17). Most of our language is composed of prefabricated expressions (Biber, Conrad and Cortes, 2004), and Wray and Perkins (2000) argue that these patterns exist in our brains as prefabricated sequences. There are many terms to describe this phenomenon. Wray (1999) uses the term *formulaic sequences.* Biber et al. (1999) talk about *lexical bundles* and Scott (1996) refers to them as *clusters*. The term formulaic sequence is a very broad term that is used to cover different sorts of multiword vocabulary items such as idioms, phrasal verbs and fixed expressions. Because there are so many terms to describe this phenomenon, *formulaic sequences* is used in this chapter as an umbrella term to include other examples or types of formulaic language such as collocations. Therefore whenever there is a reference to *formulaic sequences* it can be assumed that there is also an (indirect) reference to collocations. Most of the definitions provided in this chapter for both formulaic sequences and collocations are quite similar, having frequency of occurrence (or words that occur more with certain words than others) as one of the main characteristics of formulaic language.

Formulaic sequences, and collocations in particular, have been of increasing interest in recent years (Nattinger and DeCarrico, 1992; Wray, 1999, 2002; Van Lancker –Sidtis and Rallon, 2004; Boers, Eyckmans, Kappel, Stengers and Demecheeler, 2006; Hyland, 2008; Nekrasova, 2009; Millar, 2011; Martinez and Schmitt, 2012). Many researchers assert the importance of formulaic sequences, especially in language learning and teaching (Nesselhauf, 2003; Shin and Nation, 2008; Gardner and Davies, 2007) and second language learning (Lewis, 2000; Nesselhauf, 2003; McCarthy & O'Dell, 2005; Webb and Kagimoto, 2009; Yamashita and Jiang, 2010). What all these studies have in common is the fact that they highlight the importance of formulaic sequences in language learning and teaching, which justifies the decision to include them in the predictive model in this thesis.

Chapter 2 discussed the issue of what is involved in knowing a word. According to many, it means also understanding its collocations (Laufer 1997; Lewis, 2000; Nation, 1990, 2001). Schmitt (1999) asserts that when learning vocabulary we need to know more than just the word; we also need to understand the collocations and associations of that word. Further to the discussion in Chapter 2 on depth of knowledge, Schmitt (1999) states that measures of vocabulary may not be enough to describe a learner's vocabulary, and that other dimensions need to be added to the model, such as depth of knowledge (Read, 1993; Schmitt, 1995; Wesche and Paribakht, 1996). The need for breadth and depth of chunk knowledge is highlighted by Boers and Lindstromberg (2009). Their study is relevant to this thesis because in Study 2 an aspect of depth of vocabulary knowledge was added to the IELTS ratings predictive model. Collocational behaviour of a word (and its frequency of use) is one of the main characteristics listed by Nation (1990) - see Table 2.1 in Chapter 2- for describing depth of knowledge of a word. Learning collocations of known words can mean strengthening the depth of knowledge of those words (Webb and Kagimoto, 2009). Boers and Lindstromberg (2009) state that even if learners have a vast vocabulary, they often fail to combine the right words.

These studies support the argument that the use of collocations is seen as a feature of 'deeper' vocabulary knowledge. This is also the view that is adopted by the IELTS organisation and explains why IELTS raters, using the band descriptors, place learners who exhibit instances of correct collocational use in higher band levels.

In this study I focus on collocations (operationalised by the Martinez and Schmitt PHRASE List, 2012) as an example of formulaic language. The knowledge and use of collocations seems to be one of the main aspects of language proficiency. The use of appropriate collocations indicates a proficient user of a language. Collocations or the use of formulaic language is one of the qualities that candidates in the IELTS exam are assessed on. Therefore it seems inevitable to consider collocations as having a dependable relationship with lexical richness. The term formulaic sequences covers many other forms (phrasal verbs, idioms etc.) therefore to be consistent I use the term formulaic sequences as a generic term and collocations as a particular type of formulaic sequences. For the purposes of this study, the terms formulaic sequences and collocations are operationalised by using the phrasal expressions list of Martinez and

Schmitt (2012), which consists of the 505 most frequently used phrasal expressions in English (see below for details).

However, even though the importance of formulaic sequences has been undeniably highlighted by the above studies over the years, there is still little agreement on the characteristics, definitions and methods to be identified (Biber et al., 2004). According to Wray 'formulaic language is a puzzling phenomenon' (Wray, 1999:213). Later in the chapter the challenges of dealing with formulaic language in research are discussed.

4.2 DEFINITION: WHAT ARE FORMULAIC SEQUENCES?

The term formulaic language is a very broad term that entails many other terms - including collocations. 'Formulaic language is a term used by many researchers to refer to the large units of processing- that is, lexical units that are more than one word long' (Wray 2008:3). According to Wray (2000), formulaic language is not a single phenomenon but a collection of various phenomena depending on different data sets (native learners, L2 learners, linguistically disabled learners etc.).

First of all we need to define formulaic sequences. As already mentioned , there are many terms (Wray and Perkins, 2000, identify up to forty terms) used to describe formulaic language, such as automatic language, chunks, collocations, fixed expressions, formulae, holophrases, idioms, multiword units etc. Wray chooses the term 'formulaic sequence' which is defined as: 'a sequence, continuous or discontinuous, of words or other meaning 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, 1999:213). I consider this to be one of the more rounded and accurate definitions of formulaic sequences. Wray (2002) explains that sentences that are unconsciously recognised or processed by our brains are called formulaic. These words/phrases are not broken into smaller parts and are not processed as smaller individual chunks. They are learned, acquired and used without consideration of their literal meaning. In the introduction of her book Wray (2002) provides an excellent example of how this happens in our everyday lives. Wray reports a 1993 advertisement by Kelloggs, in which people were asked what 'Rice Krispies' were made of, and most of them were

surprised to find out it was rice. This is a brilliant example of how people acquired and used the term without realising the actual meaning of the two components. The term also includes not just single words, but larger phrases such as idioms, for example 'kick the bucket'. The language user does not treat it as 3 different words and does not capture the literal meaning. Instead it is processed as one whole unit meaning something else (not the action of kicking an actual bucket). Therefore, all the previous examples are instances of formulaic language. According to Wray (2002), there is a massive list of names/terminology regarding this phenomenon, but claims that the best term to use is 'formulaic language', due to the fact that it is a neutral term as certain other terms are associated with various researchers that were the first ones to use each term.

Even though Wray uses the term 'formulaic language' many researchers use other terms to describe it. 'Lexical phrases' is the term used by Nattinger and DeCarrico (1992), and are defined as '...chunks of language of varying length' (1992:1). Lewis (1993) made a successful attempt to bring attention to formulaic sequences or chunks with what he named the 'Lexical Approach', in which he proposed that learners need to learn chunks of the L2 language they want to learn. Wray introduced in 2008 a new term called MEU (Morpheme Equivalent Unit) which is defined as ' a word or word string, whether incomplete or including gaps for inserted variable items, that is processed like a morpheme, that is, without resource to any form-meaning matching of any sub-parts it may have' (2008:12). 'Lexical bundles' is a term first introduced by Biber and colleagues (Nekrasova, 2009). According to Biber and Conrad (1999), lexical bundles are 'three or more words that show a statistical tendency to co-occur' (1999:183). According to Wray (2002), collocations ('make a decision'), social formulas ('nice to meet you'), multiword phrases ('on the other hand'), and idioms ('shoot the breeze') all fall under the broader category of formulaic sequences (Wray (2002) cited in Zyzik, 2011). Gardner and Davies (2007) used the term 'multiword knowledge' to include a vast range of items such as idioms, phrasal verbs, fixed phrases and prefabs. However, it remains a mystery which of these should be taught and how well they should be taught (Condon and Kelly, 2002; Darwin and Gray, 1999; Nesselhauf, 2003).

The definitions given by various researchers appear to be distinguished into two categories based on either phraseology or frequency. Nesselhauf (2003) uses a

phraseological definition rather than one based on frequency or co-occurrence of words. Nesselhauf's distinction of formulaic sequences is the following:

-free combination
-collocations
-idioms

Like Nesselhauf (2003), Laufer and Waldman (2011) use the definitions: free combinations, collocations, and idioms to distinguish between the different examples of formulaic language.

Boers and Lindstromberg (2009) categorise chunks in English into various categories: strong collocations ('commit a crime'), social-routine formulae ('have a nice day'), discourse markers ('on the other hand'), compounds ('peer pressure'), idioms ('take a back seat'), standardised similes ('clear as crystal'), proverbs ('when the cat's away'), genre-typical clichés ('publish or perish'), exclamation ('you must be kidding'), and more (2009:2).

Many of the definitions are based on frequency of occurrence. A definition based on frequency is provided by Hyland, who describes formulaic sequences with the following statement: 'Essentially, these are words which follow each other more frequently than expected by chance, helping to shape text meanings and contributing to our sense of distinctiveness in a register' (Hyland, 2008:4). Biber et al. (2004) looked at lexical bundles from a frequency perspective. Altenberg (1998) was one of the first to use such an approach (frequency approach). According to Biber et al. (2004), before them only Nattinger and DeCarrico (1992) dealt with the issue of lexical bundles (they used lexical phrases) in university lectures. Cook also states that: '...actual language use is less a matter of combining abstract grammar rules with individual lexical items, and more a matter of collocation; that there are grammatically possible utterances which do not occur, and other which occur with disproportionate frequency' (Cook, 1998:57). All these studies that adopt this approach show the existence of a link between language/vocabulary use and frequency. According to Jiang and Nekrasova (2007), formulaic sequences are high frequent multiword expressions, and this is what distinguishes them from other phrases (that do not occur together with the same

frequency). What distinguishes them from idioms is the fact that their meaning can be deduced by looking at the different component words (unlike idioms). Schmitt (2010) uses formulaic language as the broader term and formulaic sequence when referring to individual examples of the phenomenon of formulaic language. Collocations falls under the term formulaic sequences and these are the conventions I follow for my own research.

4.2.1 Acquisition and Use

According to X Yu (2009), the use of formulaic sequences are in the focus of many studies in SLA research. Nattinger and DeCarrico (1992) also claim that formulaic language is the core of language acquisition.

There is sufficient evidence from various researchers (Wray, 2002; Wray and Perkins, 2000; Conklin and Schmitt, 2008) supporting the argument that formulaic language (various different types of formulaic language) are stored holistically in the brain rather than as isolated words. X Yu (2009) investigated whether using two different methods of acquiring/learning the lexical chunk 'despite the fact (that)' would produce different results. The two methods used were drilling/memorisation and teaching through explicit instructions. The subjects were Chinese first year learners of English, and were all given pre-tests and post-tests. The results showed that the group that were learning through memorisation scored higher in terms of procedural knowledge than the group that were given explicit instructions. The second group however, scored higher in terms of declarative knowledge. One possible explanation, according to X Yu (2009), is that the participants learned and memorised the phrase 'despite the fact (that)' as a chunk (not analysed in smaller parts). Wray (1999) refers to formulaic sequences and their links to aphasia. Wray explains how formulaic sequences are stored and processed by using examples from aphasic patients. Aphasic patients can often remember and recite verses of poems or songs, but cannot remember single words. Examples from aphasic patients can be an indication or proof that formulaic sequences are prefabricated and stored as single words (items).

The two previous studies (by X Yu and Wray) are relevant to the thesis as they seem to provide evidence to the argument that formulaic language is stored holistically in the

brain therefore making it easier to retrieve and use in speech or writing. This explains why collocations and formulaic language use in general are associated with language fluency and proficiency. More studies that support this claim are the following: Jiang and Nekrasova's (2007) study supports the claim by previous researchers such as Altenberg (1998) and Schmitt and Carter (2004) – also known as 'the holistic hypothesis'- that formulaic sequences are stored holistically in the brain, and therefore it is easier to access and use them. Furthermore, according to Vogel-Sosa and MacFarlane (2002), it is also assumed that multiword units (collocations, idioms etc.) are stored holistically in the mental lexicon.

According to Conklin and Schmitt (2008) and Jiang and Nekrasova (2007), formulaic sequences are processed more quickly than non-formulaic sequences. They are also processed more accurately (Jiang and Nekrasova, 2007). I believe that this could be one of the reasons that the use of formulaic language is a more common characteristic in oral speech where people do not have as much time to think about their response therefore access units that are stored holistically in the brain as a time-efficient strategy. Tremblay, Derwing, Libben and Westbury's study (2011) also showed that lexical bundles are read faster than the control sentence fragments, and this could be proof that lexical bundles are stored and processed holistically in the brain.

4.2.2 Teaching and learning formulaic sequences

Learning formulaic sequences can be very problematic (Li and Schmitt, 2009). Nonnative speakers of English will know fewer formulaic sequences than native speakers due to the fact that the latter have more exposure to the language (Wray, 2002).

The general previous belief was that learning formulaic sequences (prefabricated word sequences) would be easy to learn, but it turned out that, even for L1 learners, it is very hard to learn these sequences and this can only be achieved at a later stage of learning a language, almost as late as the teenage years (Wray, 1999; Pawley and Syder, 1983). Martinez and Murphy's study (2011) provides evidence that multiword units are hard for L2 learners to learn and understand. They state that there is a gap in research concerning vocabulary (in terms of multiword units) and reading comprehension, and prove with their study that even if learners come across a text with words from the top

2000 words in English, they struggle with the meaning (their reading comprehension is reduced significantly) if these words are presented to them in the form of multiword expression (for example, large, and, by → by and large). Kennedy (2003:467) stated that:

> 'The teaching of collocations might be expected to have a more explicit and prominent place in the language teaching curriculum. In class, teachers can draw attention to collocations not only through direct teaching but also by maximizing opportunities to acquire them through an emphasis on autonomous implicit learning activities such as reading'.

Lewis argues (1993) that learning collocations is of the same importance as language learning. Boers (2000) suggests the addition of classroom activities that enhance language learners' metaphor awareness which is linked with vocabulary acquisition, and Millar (2011) stresses the fact that there is an increasing interest in formulaic sequences and a need for them to be addressed in second language teaching and learning.

The fact that these previous studies show the difficulty learners encounter when learning or using formulaic sequences adds to the argument that these sequences are present in more advanced learners' speech or writing, supporting one of the hypotheses Study 2 is based on: that more proficient learners (learners that achieve a higher IELTS band level) should use more formulaic language than learners placed on lower level bands.

4.2.3 How to detect/find formulaic language in a text

There are some features of formulaic language that could help someone to detect it in a piece of writing (Wray, 2002). There are two ways of detecting formulaic use. The first is by conducting an experiment or handing out questionnaires targeting the language one wishes to study. The second is carried out by having a set of collected data and analysing it to reveal certain formulaic use. This method relies heavily on intuition, which is not a very scientific way of carrying out research, but is the most common. Wray (2002:20) explains that intuition is not well accepted by the scientific community and reports that Chomsky has criticised the use of intuition in experiments (even though Chomsky himself works with judgements that are based on intuition). This idea is also

supported by Schmitt (2010:65) who suggests that using intuition as a research method has many limitations and it is not always a reliable indicator of frequency. Foster (2001) has used the method of intuition to identify formulaic sequences but was criticised by Wray (2002) who discussed some inherent problems with intuition such as the fact that it can only be used with small data sets, it is inconsistent due to tiredness of individuals and there can be significant variation between judges. I believe that it is not harmful to use intuition as a method as long as it is combined with another (more reliable) method such as a computer analysis (a combination of methods is ideal).

Another means of detecting formulaic language is through frequency counts. Computers can count the occurrence of formulaic 'frames' using a set of corpora. Sinclair and Renouf (1988:151) state the importance of frequency counts but report that it cannot be the only aspect that is important in detecting formulaic sequences. Nevertheless, it is very hard to detect formulaicity because of the difficulty deciding on a single definition that covers all aspects of formulaic sequences. In order to locate formulaic sequences in a text we could decide on an 'exclusive' definition that may exclude some forms of formulaic use (Wray, 2002). This has implications on developing automated ways of predicting teacher ratings as all variables entered in the model need to be pre-defined by the researcher.

4.3 DIFFERENT TYPES OF FORMULAIC SEQUENCES

Wray (2000) suggests that most researchers who attempt to categorise formulaic language do so by trying to separate form and function. A short description and discussion of idioms and phrasal verbs follows in this section but is beyond the scope of this study to expand on this discussion. I give emphasis to collocations (in a separate section), as an example of formulaic sequences, which is the focus of this study.

4.3.1 Idioms

Although idioms are not part of the main focus in this study, there is a description of idioms in this sub-section. Collocations, idioms and metaphors are all considered to be subcategories of formulaic language (Wray, 2002; Simpson and Mendis, 2003). Learning and using idioms is very problematic for all types of learners, especially L2 learners (Cooper, 1999). A learner that masters idioms in a foreign language is

considered to be fluent in that language according to many researchers (Simpson and Mendis, 2003; Schmitt, 2000; Wray, 2000). The usage of idioms is a very useful aspect for language learners of English (Liu, 2003).

Defining idioms however, can be very problematic (Simpson and Mendis, 2003). According to Wray (2002), idioms are considered to be the main representative example of formulaic sequences for many researchers. Among the many definitions provided over the years were the following: Wood (1986:2) describes an idiom as 'a complex expression which is wholly non-compositional in meaning and wholly non-productive in form'. Nattinger and DeCarrico (1992:33) describe idioms as 'complex bits of frozen syntax, whose meanings cannot be derived from the meaning of their constituents, that is, whose meanings are more than simply the sum of their individual parts'. A similar definition is provided by Cooper (1999) who states that: 'An idiom is an expression whose meaning cannot always be really derived from the usual meaning of its constituent elements. It is hard to tell from the literal meaning of the individual words, for example, that *to kick the bucket* or *to bite the dust* means to die'(1999:233). One has to add that there is also a literal meaning of these expressions. People can kick a real bucket and then this expression does not mean to die of course. Simpson and Mendis (2003) used the following definition for their research of idioms in academic speech: '...an idiom is a group of words that occur in a more or less fixed phrase and whose overall meaning cannot be predicted by analyzing the meanings of its constituent parts' (2003:423).

Three criteria which were previously used by other researchers (Fernando,1996, McCarthy, 1998, and Moon,1998) were also used to define idioms: *Compositeness or fixedness* which describes the non-ability to replace or substitute any of the specific individual words in an idiom (for example, *off the deep end*), *institutionalisation* which describes the acceptance of the expression by a wider community, and *semantic opacity,* which means that one cannot guess the meaning of the idiom by analysing its individual parts because it would not make sense (Simpson and Mendis, 2003). They compiled a list of 32 frequent idioms in academic speech using a corpus-based study. Grant and Bauer (2004) made an attempt to try and redefine idioms. They claimed that the existing definitions were not specific and adequate enough so they proposed a more restrictive definition of idioms in the form of a test which divides multiword units into

three categories: core idioms, figuratives and ONCEs (one non-compositional element). A core idiom is defined as an idiom that is non-compositional and non-figurative and there is more than one element in the MWU that is non-compositional. Figuratives are idioms that use figurative language such as metaphors. An idiom is considered an ONCE when only one word of the MWU is found to be a non-compositional, necessary part (Grant and Bauer, 2004:53).

According to some researchers (Fernando, 1996; Moon, 1998; Nesselhauf, 2003) idioms should not be considered as collocations because collocations are fairly transparent, but others disagree and claim that idioms should be considered collocations (Palmer, 1933; Wouden, 1997). In this study I align with the latter as I agree that idioms can go under collocations, when loosely defined is a grouping of words that form a phrase or clause (a very generic term). However, even when problems with defining idioms are addressed, what idioms should be learned or taught in order for learners to become fluent is still open to further research (Simpson and Mendis, 2003).

4.3.2 Phrasal Verbs

Phrasal verbs is another example of formulaic language but since it is not in the scope of this study to investigate phrasal verbs only a short description is provided here. Gardner and Davies' (2007:341) definition of phrasal verbs is provided below:

> '...Any two-part verb consisting of a lexical verb (LV) proper followed by an adverbial particle (tagged as AVP) that is either contiguous (adjacent) to that verb or non-contiguous (i.e., separated by one or more intervening words)'.

This definition was chosen by the researchers instead of Biber et al.'s (2004) because of its simplicity, as it only entails one syntactic criterion: 'a verb plus an AVP'. Biber et al.'s definition involves an extra semantic component. Liu (2011) suggests that phrasal verbs are extremely frequent yet very hard for L2 learners to understand and use. They are mostly used by advanced learners of a foreign language. Therefore it can be assumed that they are an indication of proficiency (as most examples/types of formulaic language).

4.4 COLLOCATIONS

4.4.1 Definition of collocations

In this section various collocations definitions are presented and discussed. One of the definitions provided by Möbarg describes collocational view of language: 'the view that words are not isolated, individual units, with no other potential for combination with other words than a formal tagging, but that, on the contrary, they tend to appear predictably together with certain other words' (Möbarg, 1997:204). According to Möbarg (1997), this view has been adopted widely in recent years mainly due to technology which helps researchers to analyse large amounts of data (corpora) and find statistical tendencies.

Collocations (as well as formulaic sequences in general- see definition of formulaic sequences in section above) can be defined from a statistical standpoint- 'frequency of co-occurrence of two lexical items within a given span' (Webb and Kagimoto, 2009:59; Greenbaum, 1974; Partington, 1998; Sinclair, 1991), or a phraseological standpoint (Cowie, 1994; Nesselhauf, 2003). The above first term (statistical point of view) has been widely accepted by corpus linguists such as Halliday (1966), Sinclair (1991) and McEnery and Wilson (2001). This is also supported by Walker (2011) who claims that there are two main categories of definitions: the lexical approach to collocations, and the frequency approach. Researchers that use the lexical approach seem to choose collocations in terms of lexical criteria such as fixedness or opacity (Carter, 1987; Cowie, 1998; Howarth, 1996, 1998), whereas researchers that use the second approach seem to pick collocations as words that co-occur together (Moon, 1998; Nesselhauf, 2003, 2005; Sinclair, 1991). The second approach is adopted in my study (Study 2) since the list used for the analysis of formulaic is based on frequency of occurrence (505 most frequent phrasal expressions).

Carter (1988:163) provides the following definition: 'A collocation is an aspect of lexical cohesion which embraces a 'relationship' between lexical items that regularly co-occur'. A very simple definition of collocation is provided by Aghbar (1990) and states that collocations are stored as two combined words (in the memory of native speakers) and are frequently found together in oral and written speech. Nattinger and

DeCarrico (1992) provide the following definition of collocations: '...describe specific lexical items and the frequency with which these items occur with other lexical items ...' (1992:20). Another definition for collocations is provided by Shin and Nation (2008:341): 'Collocation is used to refer to a group of two or more words that occur frequently together, and it is not restricted to two or three word sequences.' According to Shin and Nation, collocations consist of two parts: the main/focal word, and the collocate. The chosen definition for people that are dealing with language teaching and lexicography is the habitual combinations of words. Language educators and researchers do not agree on what word combinations can go under the header of 'collocations' (Liu, 2010). Wouden (1997:53) also adds to the discussion by stating that: 'what goes under the header of 'collocation' is very heterogeneous'. A definition given by corpus linguistics is the following: the co-occurrence of lexical items 'with greater than random probability' (Hoey, 1991:6-7). Shin and Nation (2008) state that even though Lewis (1993) work was very influential, he did not have a clear classification of what a multi-word unit is. Therefore, the researchers decided to try to clearly define collocations and find the most frequent ones so they could be used in elementary language teaching classes. Most teachers (language educators) seem to be in agreement about the following definition: collocations are combined words with restricted cultural variation and are not just free word combinations or idioms (Liu, 2010). Even choosing the above definition, it is still hard to decide which word combinations are collocations.

Collocations are considered by many researchers and educators to be arbitrary (Benson, 1989; Smadja and McKeown, 1991; Lewis, 2002; Nesselhauf, 2003). According to Liu (2010), collocations are not arbitrary, but there is semantic motivation behind each choice; therefore collocations should not only be taught as fixed chunks. According to Yamashita and Jiang (2010:649), collocations are different from formulaic sequences because they are 'looser combinations of words than formulaic sequences, in the sense that a component word in a collocation may collocate frequently with many other words to form other collocations'.

In the present study the term formulaic sequences is used as a generic term which covers collocations as they are an example of formulaic sequences. Following Nation's example 'the term collocation will be used to loosely describe any generally accepted

grouping of words into phrases or clauses' (Nation, 2001:317). Therefore all vocabulary items (phrases) that comprise Martinez and Schmitt's (2012) list are considered collocations.

According to Schoonen and Verhallen (2008), the use of collocations is a dimension of depth of lexical knowledge (others are meaning, grammatical category, derivations, pragmatic and sociolinguistic value – Nation, 2001). Indeed, collocations are considered to be an aspect of depth of knowledge (Beglar and Hunt, 1999). Furthermore, collocational properties (and frequency) fall under the 'depth' category according to Qian (1999). Refer back to Chapter 2 for a detailed discussion on dimensions of lexical knowledge-depth of knowledge.

4.4.2 Collocations- Acquisition and Use

There are various studies regarding the acquisition and use of collocations. Bahns and Eldaw's (1993) study showed that 48% of all the errors in EFL learners' productive knowledge were collocational errors (even though the collocations percentage of lexical words was 23%). They concluded that advanced learners face a problem with the use of collocations and suggested that the learning of collocations is not increased in parallel with general vocabulary knowledge. Farghal and Obiedat's (1995) study involved Arabic EFL students. They were tested on their written productive knowledge of adjective & noun, and noun & noun English collocations. The researchers found that these learners faced a major problem with coping with collocations because they were not aware of the existence of collocations. Therefore, the students seemed to either literally translate from their L1 or use synonymy or paraphrase to cope with collocation. Fan (2009) argues that various researchers, for example Pawley and Syder (1983), Hunston and Francis (2000), and Wray (2002) all claim that the use of collocations is an important aspect of L2 acquisition and use. When a language learner uses collocations it makes their speech (and writing) sound more native-like and fluent. Nevertheless, Fan (2009) continues to say that no matter how proficient learners are, they always have problems with the use of collocations (Fan, 1991; Biskup, 1992; Farghal and Obiedat, 1995; Nesselhauf, 2003; Ellis, Simpson-Vlach and Maynard, 2008; Laufer and Waldman, 2011). Webb and Kagimoto (2009) also stated that the use

of collocations can be problematic for L2 learners. Fan (2009) states that one quarter of the collocations used by the learners are not correctly used (contain mistakes).

Walker (2011:291) proposes, in contrary to previous research which claims that collocations are arbitrary, that there is an explanation for the collocational choices used by speakers, and he suggests that '... if the learner is encouraged to look for an explanation, it makes the process of learning collocations more memorable'.

One of the problems with collocations is the fact that when trying to express a single meaning, two different words can be used in different combinations. Fan (2009) uses the example by Halliday (1966) of the words 'strong' and 'powerful', where they have the same meaning but when talking about a car we usually use the word 'powerful' but when talking about tea we normally say 'strong' tea. Second language learners may use grammatically correct sentences but not idiomatic sentences- choices used by native speakers (Wray 1999).

Another problem with collocations is L1 interference (Schmitt, 1999; Nesselhauf, 2003; Wray, 2002; Fan, 2009; Boers and Lindstromberg, 2009) because in the learner's L1 the word may have a different collocate than in their L2. Wolter and Gyllstad's (2011) research results support the previous statement. Their results showed that: 'the L1 may have considerable influence on the development of L2 collocational knowledge' (Wolter and Gyllstad, 2011:430). L1 influence may be helpful at times but can be inhibitory at others (Wolter, 2006). According to Nesselhauf (2003), L1 can influence the production of collocations. Therefore, learners' L1 should not be abandoned in language teaching. Nesselhauf (2005) stated that L1 influence is more than evident in her research, as this is exhibited in more than 50 per cent of the collocation errors learners made. Fan (2009) reports that this is indicated in her research also, as the L2 learners were using a variety of collocations which did not seem to be acceptable in English but were direct translations from Chinese (they were acceptable collocations in Chinese, their L1) . On the contrary, a quite recent study by Yamashita and Jiang (2010:647) regarding the L1 influence on the acquisition of L2 collocations shows that 'once stored in memory, L2 collocations are processed independently of L1'. Since in my study (Study 2), like Farghal and Obiedat's (1995) study, Arabic EFL learners were

involved the use of collocations was not high. The researchers suggested that the use of collocations was problematic for Arabic learners as most of them were not even aware of the existence of collocations. Therefore, the previous studies, which highlight the difficulty acquiring and using collocation, are relevant to my study and may at a later stage provide possible explanations or interpretations to my findings.

Fan (2009) also reports that the use of collocations is affected by the size of the learner's vocabulary and grammar. This was illustrated in the L2 learners' essays as they had not successfully attempted to describe a picture due to lack of vocabulary. Fan (2009) concluded by stating that L2 learners proved to have a smaller vocabulary due to the smaller number of collocations used in their essays. This study provides evidence to the argument that there is a link between vocabulary size and the use of collocations. This finding seems to be supported by literature as Kaszubski (2000), Lorenz (1999), Granger (1998) and Chui (2006) came to a similar conclusion. Thus, according to Fan, the more words a learner knows (the larger the vocabulary), the more collocations they are expected to use. Laufer and Waldman's study (2011) showed that nonnative speakers use fewer collocations than native speakers. However, the number of collocations (verb-noun collocations) used by nonnative speakers increases at an advanced level. Language learners seem to know and use fewer types of collocations than native speakers, but they tend to overuse the ones they know (Hasselgren, 1994, Cobb, 2003). Learners overuse and over rely on some collocations or structures they feel comfortable with; Hasselgren (1994) calls these 'lexical teddy bears'. The learners end up sounding strange because of this overuse (Cobb, 2003). Dechert (1983) states that prefabricated chunks are used by learners as 'islands of safety' (Dechert 1983 cited in Boers and Lindstromberg, 2009). Siyanova and Schmitt (2008) also suggest that L2 learners use many collocations, but not all of them are appropriately used. Webb and Kagimoto (2011) suggest that L2 learners learn more collocations with an increased number of collocates per node word (the greater the number of collocates per node word, the more easily the collocations are learned). Node words are the base words used for the different collocations. For example, for the collocations *good laugh, good reason* and *good behaviour,* the node word is *good* (Webb and Kagimoto, 2011:7). Some of the methods that have been used to research the acquisition and use of collocations by L2 learners are error analysis, collocations elicitation, and analysis of learners' corpora (Laufer and Waldman, 2011). For Study 2 I am taking on two of these

methods: collocations elicitation (using the PHRASE List by Martinez and Schmitt, 2012) and analysis of learners' corpora.

4.4.3 Collocations and Frequency Factor

This section highlights the importance of the frequency factor regarding collocations and formulaic language in general. The frequency factor was previously mentioned in an attempt to provide possible definitions for formulaic sequences and collocations in particular. Most researchers approach the terms from two perspectives: the phraseological or the frequency perspective. The frequency factor is very important in collocations research and vocabulary research in general. It is quite obvious by looking at previous sections (definitions of formulaic language and collocations) how important the frequency factor is in defining formulaic language (and collocation in particular) as most of them mention frequency or frequency of occurrence etc.

Wray (2002:25) stressed that frequency is a salient factor in the identification of formulaic sequences. According to Shin and Nation (2008), frequency is not the only factor determining collocations, but it could be a good start for learning spoken English and improving fluency. Liu (2003) also suggested that frequency alone is not enough to determine what is important for language teaching. Millar (2011) states that frequency is a well-established factor that affects the use and storage of formulaic sequences. However, formulaicity cannot be defined in terms of frequency alone (Wray, 1999). Wray and Perkins (2000:6-7) also suggest that frequency is not the only important factor in terms of the use of formulaic sequences. On the other hand, Vogel-Sosa and MacFarlane (2002) suggest that collocation frequency is the main factor that determines lexical storage. My view on this point is that frequency seems to play an important role, especially in the identification of formulaic sequences (as most lists are based on frequency). Frequency of use certainly has implications on the usefulness of specific collocations for teaching and learning.

4.5 WORD LISTS AND ACADEMIC CORPORA

In the previous section the importance of frequency in defining collocations (and formulaic language in general) was highlighted which leads to the present section and the importance of word lists and academic corpora (which are used in defining

formulaic language). Therefore a description of some of the lists used is provided below.

As already mentioned in previous chapters, West's GSL (1953) is a list containing the most useful 2000 word families in English and was compiled from a 5 million word corpus. Even though the GSL has been under criticism for various reasons, such as its size (Engels, 1968), age (Richards, 1974), and the need for an update (Hwang, 1989), the GSL has remained the most commonly used list until very recently before Brezina and Gablasova, (2013) proposed a new version of it. Its coverage of fiction texts, nonfiction texts and the academic corpus is very large (Coxhead, 2000). In 2000 Coxhead proposed a new list, the AWL (Academic Word List). The AWL represents an academic extension of the GSL. The AWL is a list of 570 word families and was developed from a written academic corpus of 35 million running words (Coxhead, 2011:355). The 2000 most frequent word families of West's GSL (1953) are excluded from the count because the list focuses on academic vocabulary not general English. The three aspects taken into account for the list's selection process were frequency, range, and uniformity. Hyland and Tse (2007) are very critical of the use of the AWL. They criticise the wide use of the AWL, as stated in their study:

> 'The findings suggest that although the AWL covers 10.6% of the corpus, individual lexical items on the list often occur and behave in different ways across disciplines in terms of range, frequency, collocation, and meaning. This result suggests that the AWL might not be as general as it was intended to be and, more importantly, questions the widely held assumption that students need a single core vocabulary for academic study' (Hyland and Tse, 2007:235).

The list by Coxhead examined the words outside West's (1953) first 2000 most frequent words. This list can be used as the basis for research regarding academic vocabulary (Coxhead, 2000).

The University Word List (Xue and Nation, 1984) consists of 836 word families and the overlap with AWL is 51%, which means there are 435 word families that exist in both lists. Even though the AWL is a smaller list it covers more subject areas than the UWL and has a higher coverage of academic texts (Coxhead, 2000). This list can be used as a basis for teaching vocabulary for EAP courses. Liu (2003) developed four

lists of the most frequently used idioms in oral American English. Gardner and Davies (2007) attempted to create a list with the most frequent phrasal verbs using the 100-milion-word British National Corpus (BNC).Various lists have been formed to help serve as the General Service List did (Martinez and Schmitt, 2010; Shin and Nation, 2008; Simpson-Vlach and Ellis, 2010).

Due to the recent increase in interest concerning formulaic sequences, Simpson-Vlach and C. Ellis (2010) decided to produce a list comparable with Coxhead's (2000) Academic Word List (AWL). They created the Academic Formulas List (AFL) which comprises of formulaic sequences that are common/frequent in academic speech and writing. Simpson-Vlach and Ellis (2010) made an attempt to compile a list of the most useful formulaic sequences in Academic English. They used both quantitative and qualitative methods to compile the list, and after combining the two could predict which sequences were worth teaching. The ones that were found 'Formula Teaching Worth' (FTW) were those that were put first on the Academic Formulas List (AFL). Even though they used similar methodologies with Martinez and Schmitt (2012), the main difference between the two lists is the subjectivity issue in terms of selecting the phrases to be included in the list. The raters' judgement was not used to influence the selection of the phrases, but was only used in the model (multiple regression) alone. It was used to inform the multiple regression but 'did not directly influence the selection of items' (Martinez and Schmitt, 2012:306). Whereas in Martinez and Schmitt's (2012) list, the authors' subjectivity played a major role for the selection of items. This is an advantage of their list because as Nation (2001:56) suggested 'Studies on collocation which have relied solely on computing procedures have yielded results which are not very useful.' Even though the issue of subjectivity can be seen as a disadvantage of their list, up to date there is no computer application that can identify, recognise and determine the formulaic sequences that most speakers use (Fatahipour, 2012). One of the differences between the two lists is the fact that for the AFL the chosen items were not ranked by how commonly they were found in discourse which is the case for most lists, including the PHRASE List (Martinez and Schmitt, 2012:306). The second list was used in a study by Fatahipour (2012) who investigated the validity of lexical diversity measures and their correlation with the use of formulaic sequences. Therefore, the same list is used in the present study in order to examine whether the analysis produces the same results.

Another list of formulaic language and in particular collocations was created by Shin and Nation (2008). They attempted to compile a list of the most frequent collocations in oral English based on six criteria. They included semantics as one of their criteria. Even though their methodology is similar to Martinez and Schmitt's, they did not include frequency or semantic transparency as their criteria, neither the degree of usefulness. Thus, even though some items are common in both lists, such as 'a bit' and 'as well as', there are other more transparent phrases that could not be included in this list, such as 'this year' or 'very good' (Martinez and Schmitt, 2012). Shin and Nation (2008) proposed a list of the most frequent collocations in oral English and suggested that these should be taught in English speaking courses (at elementary level). There are some limitations to the list because of the corpus used for this study.

One of the most recent attempts to include formulaic sequences in pedagogic materials such as EFL textbooks and vocabulary tests was made by Martinez and Schmitt (2012). They compiled the PHRASal Expressions List (PHRASE List), which presents the 505 most frequent non-transparent multiword expressions in English (Appendix 8). Martinez and Schmitt (2010) created a list with the 505 most common multiword expressions using the BNC (British National Corpus). They chose frequency and noncompositionality as their main criteria (Martinez and Murphy 2011). Martinez and Schmitt (2012) state wordlists can be very useful for pedagogical purposes. The GSL and the AWL are great, but have one major drawback, which is the fact that they only concentrate on individual words. Therefore, there was a need for a list to be compiled to address this gap, and a need for a list that went beyond the individual word level, such as the GSL (West, 1953) and the AWL (Coxhead,2000). This is exactly what Martinez and Schmitt's (2012) list was for. The list is mainly based on frequency as it was created to serve pedagogic purposes and according to various researchers such as Leech (2001) and Nation (2001), 'frequency of occurrence is one of the best indicators of usefulness of individual words in general English' (Martinez and Schmitt, 2012:4). The researchers did not want to base the list on frequency alone, therefore they chose phrases 'that conveyed a discrete, identifiable meaning or function' (2012:5) as one of their selection criteria. Another criterion was the transparency (compositionality) of the phrases. They (2012) chose phrases whose meaning could not be easily understood (by their constituent parts). They chose less compositional phrases to avoid the issue of learners trying to interpret the meaning (phrases that are harder to learn). To sum up,

their criteria was 'high frequency, meaningfulness, and relative non-compositionality' (Martinez and Schmitt, 2012:6), which make the list comparable with single word lists. There were core criteria and auxiliary criteria for the items selection. The core criteria were the following: the expression had to be a Morpheme Equivalent Unit (MEU), not semantically transparent, and not be deceptively transparent. The auxiliary criteria were: the expression may have a one-word equivalent either in English or another language, the L1 could negatively influence interpretation, and the meaning or opacity of the word should not change because of grammar. The criteria were not cumulative. There are also limitations to the list which have been raised in recent studies (Fatahipour, 2012). The list does not claim to cover all formulaic sequences, but only a 'limited subset of formulaic language' (2012:6). Fatahipour (2012) found from his analysis that certain common collocations were missing from the list. Furthermore, the researchers decided to name their category of formulaic language (this subset) 'phrasal expressions'. Their definition is the following:

> 'A phrasal expression is hence 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 it is easily discernible by decoding the individual words alone' (Martinez and Schmitt, 2012:6).

This is quite confusing because yet another term is added to the already very long terminology list to describe the formulaic language phenomenon. Due to this fact and to be consistent and have comparable results I put the examples from their list under the more general term of collocations.

Möbarg (1997) criticises vocabulary lists by arguing that any list, no matter how sophisticated, is not the real vocabulary of a language, but rather a sample of it. This is a criticism I would align with, the stance taken however is that attempts to produce more sophisticated lists can improve applications and tests. Vocabulary lists will always be useful since it is extremely hard and almost impossible for a researcher to be present and analyse learners' real vocabulary use. Therefore researchers will always need to rely on either smaller samples of data or vocabulary lists.

4.6 FORMULAIC LANGUAGE (COLLOCATIONS) AND PROFICIENCY

Vocabulary is one of the main aspects of L2 proficiency (Schmitt, 1999; Hsu 2007). However, Schmitt (1999) claims that it is not enough to just know a word, learners need to know all the collocations and associations of that word (need to go into more depth). 'Good chunk knowledge does contribute to proficiency in L2 as well as in L1' (Boers and Lindstromberg, 2009:38).

Wray (2002) states that a high level of proficiency in a non-native language indicates that the user can not only use single words, but also knows how to combine them and can identify and learn idiomatic and native-like sentences. Pawley and Syder (1983) also claim that a high-level language user should be able to select sentences that sound more native-like, instead of other sentences that could be used (they are grammatically correct) but are not normally used by native speakers. Schmitt (1999) states that knowing a word should mean more than just recognising it. Learners should know all the associations and collocations of that word.

Schmitt (1999) also states that it is not clear whether vocabulary's predictive validation can be linked with a learner's general language performance because other factors (L1) may interfere. Wray (1999:213) proposes that: 'Formulaic language offers processing benefits to speakers and hearers, by providing a short cut to production and comprehension'. This happens because the sequence is stored as a single big word (Ellis, 1996:111) and with an associated holistic meaning.

Proficient learners are expected to have more developed semantic networks in the L2 mental lexicon (Wolter, 2002:315). Wolter (2002) wanted to test if a WAT (Word Association Test) would be a suitable tool to test learners' proficiency. The test, which comprised of 20 verbs which were taken from the Edinburgh Associative Thesaurus EAT (Kiss, Armstrong, Milroy and Piper, 1973), requested 3 responses from the learner. C-tests were used to assess proficiency and the scores were correlated with the WAT scores using a Pearson-*r* correlation test. The results were statistically significant but the correlations were moderate. Even though at that point the test did not seem to be an ideal medium to measure proficiency, the researcher believes that it could be (in the future) if the test is carefully constructed.

According to Shin and Nation (2008), learning collocations is essential if learners want to improve their fluency and sound more native-like. This is also supported by Pawley and Syder (1983), who also state that even though learners often produce grammatically correct sentences, they do not sound like native speakers because they do not use correct collocations (collocates). Gardner and Davies (2007) and many others (Nattinger and DeCarrico, 1992; Moon, 1997; Nesselhauf, 2003; Wray, 2000, 2002; Schmitt, 2004; Hyland, 2008; Li and Schmitt, 2009) also argue that multiword knowledge is an important aspect of native like fluency. Al-Zahrani (1998) found a strong correlation between the participants' use of collocations and their language proficiency (measured by the TOEFL test). Advanced learners seem to make more mistakes related to misuse of collocations than less proficient users (Nesselhauf, 2003; Boers and Lindstromberg, 2009). Failure to use native-like formulaic sequences can make one's writing seem non-native (Li and Schmitt, 2009). Simpson-Vlach and C. Ellis (2010:487) also highlight the importance of formulaic sequences for L2 fluency and state that: 'Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing'.

Millar (2011) also stresses the existence of the relationship between the use of formulaic language and proficiency. The author states that the way formulaic language is stored and accessed in our mental lexicon '...enables the maintenance of fluency' (Millar, 2011:131). Laufer and Waldman (2011) suggest a relationship between receptive knowledge of collocations and general vocabulary knowledge. This was also suggested by Bonk (2000), whose research results showed a high correlation between general English proficiency and collocation proficiency.

All the studies in this section provide evidence to the argument that there is a link between the existence (use) of formulaic language in a learner's speech or writing and proficiency. This argument was the basis for the formation of one of the main hypotheses of Study 2. The fact that higher level band learners (more proficient learners) would use more (have more examples of) formulaic language in their essays.

4.6.1 Formulaic language- Findings from previous studies

Formulaic language comprises one third to one-half of all language use according to various researchers (Howarth, 1998a; Erman and Warren, 2000; Foster, 2001). A large

proportion of written and oral text consists of lexical chunks (Boers and Lindstromberg, 2009). Biber et al.'s (1999) findings showed that lexical bundles constituted around 28 % of conversational text and 20 % of academic text. In Van Lancker-Sidtis and Rallon's study (2004) it was found that formulaic expressions constituted almost 25% of the phrases in the text examined (screenplay 'Some like it Hot'). The authors state that '...formulaic expressions constitute a significant proportion of discourse' (2004:217). They counted all the types (unique expressions) and tokens (repeated occurrences). Most of the occurrences of formulaic expressions in their analysis were classified as 'speech formulas' and the most frequent were the two-word expressions.

4.6.2 Formulaic language & Writing

The relation between formulaic language and writing is very important for my research as for my analysis only essays (written data) are used. Li and Schmitt (2009) state that formulaic sequences play an important role in L2 writing because '...as a result of their frequent use, such [sequences] become defining markers of fluent writing and are important for the development of writing that fits the expectations of readers in academia' (2009:86). Howarth (1998a) stated that 31-40% of the 238,000 words of academic writing that was investigated were collocations and idioms.

A study by Erman and Warren (2000) showed that 52.3% of the written data that was investigated were formulaic sequences. Some sequences such as 'as a result' are considered essential in academic writing (Coxhead and Byrd, 2007; Hyland, 2008), and lack of those in an academic text can be an indication of an unsuccessful L2 academic writer. Following the suggestions of the above studies, it is expected that in Study 2 learners' essays will comprise a percentage of formulaic language.

4.6.3 Formulaic language & Oral data (speech)

This section discusses the results from studies that investigated formulaic language using oral data. Even though oral data were not used in Study 2, this section was included to add to the discussions of the importance and presence of formulaic language in language production in general and to the discussion of the link between the use of formulaic language and language proficiency.

Hsu and Chiu (2008) tested the relationship between lexical collocations and speaking proficiency using a sample of Taiwanese EFL learners. After collecting data from 56 university EFL learners, they found that there was a relation between the learners' knowledge of lexical collocations and speaking proficiency.

Shin and Nation (2008) suggested that collocations appear more in oral speech (Shin, 2007). They did not use word families for the count but each main word was treated as a word type because it may have had a direct collocate. Their findings show that a very large number of collocations use the first 1000 words of English. Frequency played a major role in the number of collocations. 'The more frequent the pivot word, the greater the number of collocates' (Shin and Nation, 2008:343). Another finding was that short collocations are more frequent than larger ones.

Sorhuss (1977) reported 20% of formulaic expressions in Canadian oral speech.

Research by Foster (2001) showed that raters came to the conclusion that 32.3% of speech was comprised of formulaic sequences. This finding is also supported by Millar (2011) who writes: 'what corpus studies do show is that fixed or semi-fixed word combinations make up a substantial proportion of natural language use' (2011:131).

4.6.4 Formulaic language and teacher ratings

Research has also been conducted regarding collocations and language assessment. Read (2000) seemed to be puzzled by how multiword items could be used for language assessment.

Hsu (2007) investigated the relationship between lexical collocations and the online writing scores of Taiwanese College English Majors and non-English Majors. The results showed a significant correlation between the frequency and variety of lexical collocations and writing scores. It was also found that the students' essay length (total words of essays) correlated significantly with the students' online writing scores.

According to Boers and Lindstromberg (2009), L2 learners benefit from chunk knowledge. Boers et al. (2006) investigated the relationship between formulaic sequences and oral proficiency and found that the formulaic sequences count correlated well with the oral proficiency ratings. Their research showed that formulaic sequences seem to facilitate fluency due to the fact that students' fluency correlated with the number of chunks they used (correlation coefficient was .045). Boers and Lindstromberg (2009:37) reported in their book:

> 'Boers et al. (2006) and Stengers (2009) both found positive correlations between the number of lexical phrases used by the students and the scores they were awarded by blind judges for the parameter of 'range of expression'(that is lexical richness and syntactic complexity)'.

This is also reported in Millar (2011) who reports that Boers et al., study (2006) 'showed that the number of formulaic sequences used correlates well with oral proficiency ratings' (Millar, 2011:135).

Pike (1979) found a strong association between TOEFL vocabulary scores and reading items. Hill (1999:5) states that: 'students with good ideas often lose marks because they don't know the four or five most important collocations of a key word that is central to what they are writing about'. Ohlorgge (2009) came to the conclusion (after examining 170 EFL compositions) that the people with higher scores were using more formulaic sequences than those who scored lower. In addition, Lewis (2008, cited in Martinez and Schmitt, 2012) analysed EFL university compositions in Sweden and also found that there is a correlation (relationship) between the use of formulaic sequences and the grades awarded to the learners for their essays (the more sequences, the higher the grades). My study (Study 2) aims to add to the findings from previous studies and provide additional evidence to the link between formulaic language and teacher ratings.

4.7 METHODOLOGICAL PROBLEMS WHEN CONDUCTING RESEARCH WITH FORMULAIC LANGUAGE

4.7.1 Problems with collecting/eliciting data

Schmitt (1999) faced problems when measuring collocations due to the fact that he was using an experimental procedure which was probably not the best method of capturing collocational knowledge. Shin and Nation (2008) point to the disadvantages of using a

corpus for their study due to the corpus nature which is limited, as both Cook (1998) and Widdowson (2000) highlighted. Cook (1998) states that computer corpora provide information about production but not about reception. Liu (2003) also indicated the limitations of using a corpus.

4.7.2 Problems with definitions and operationalisation of terms

The definition of collocation is problematic (Gardner, 2007). Fan (2009) reports on how problematic the definition of the word 'collocation' is. It is mentioned that this word has been given many names, including multi-word units, and argues that the only agreed upon explanation, according to Nesselhauf (2005), is the fact that collocations share the following characteristic: 'some kind of syntagmatic relation of words'.

Schmitt (2010) rightly argues that there are problems with formulaic language research and many incomparable studies due to different counting methodologies. Various researchers in corpus studies come to different conclusions regarding formulaicity, and this is due to the fact that people count different aspects (some count tokens and other types) (Wray, 1999). Take for instance the example of Altenberg (1998) and Moon (1998). Altenberg came to the conclusion that around 80% of adults' speech in their L1 can be formulaic, whereas Moon (1998) suggests that only a small percentage (around 4 to 5%) of the 18 million-word corpus used was found to be formulaic. This divergence was due to the fact that the two researchers were counting different things (Altenberg counted tokens, not types). Liu (2010) examined the Corpus of Contemporary American English and the British National Corpus and the studies of Biber et al. (1999) and Gardner and Davies' (2007) to compile a list of the most common Phrasal Verbs in American and British English.

The main conclusion from various studies is the fact that any discourse consists of many formulaic sequences (Conklin and Schmitt, 2008). According to Conklin and Schmitt (2008), even though most of the studies mentioned concern only English, there have been various studies that indicate that formulaicity exists in other languages such as Russian (Cowie, 1998), French (Arnaud and Savignon, 1997) and Swedish (Bolander, 1989).

Even though there has been increasing interest in recent years on collocations (definitions, use, learning and teaching), there are still many research gaps that need to be addressed (Liu, 2010).

4.8 RATIONALE AND OPERATIONALISATION OF FORMULAIC SEQUENCES

Formulaic sequences (in this case collocations) are included here to try to improve the predictive validity of the IELTS ratings model, as they play a major role in the IELTS rating scales as part of the Lexical Resource component. A study by Brown (2003) which qualitatively analysed examiners' comments regarding their decisions on Lexical resource revealed that many examiners commented on appropriateness or correctness of collocations, and use of idiomatic or colloquial terms.

Therefore, the focus of this study (Study 2) will be the use of formulaic sequences and how this affects the model of predicting IELTS Writing scores (and more specifically the vocabulary ratings). One of the main studies regarding lexical richness measures and their relationship with formulaic language use is by Fatahipour (2012). He investigated the validity of various measures of lexical richness and then counted all instances of formulaic language to check if there was a strong and significant relationship between them. His results revealed that there was a modest (but not significant) correlation between measures of lexical richness and the use of formulaic sequences. However, there was also a modest correlation between the number of formulaic sequences and language ability scores. 'The presence of such relationship is an indication that the inclusion of FSs as a dimension of language ability is worth exploring' (Fatahipour, 2012:224). He suggests that the presence of formulaic sequences should not be ignored in any analysis between lexical richness measures and language ability. Therefore, I sought to investigate this relationship in my set of data, using a similar method.

Millar (2011) operationalises formulaicity through collocations. This thesis follows Millar's example and operationalises formulaicity by counting collocations using the PHRASE List (Martinez and Schmitt, 2012) of the 505 most common phrasal expressions in English. This is the same list that was used in Fatahipour's study for operationalisation of formulaic sequences. In the IELTS band descriptors collocations are mentioned as one of the more advanced level features. Therefore, because I am

trying to predict the IELTS ratings it is more logical to count collocations than anything else (such as idioms, for example).

Research (Bonk, 2000; Laufer and Waldman, 2011) suggests that there is relationship between proficiency and the use of collocations. Therefore, if the number of collocations is used as another variable in my model it could improve its predictive validity.

The method I follow is the extraction of collocations from learner corpora to explore learner knowledge. According to Siyanova and Schmitt (2008), this methodology was used by various researchers such as Granger (1998), Howarth (1998) and Nesselhauf (2003, 2004).

The Martinez and Schmitt PHRASE list was chosen for the operationalisation of formulaic sequences here because, as the authors suggest, the list could provide a basis for vocabulary tests, and it seemed suitable for this research as it deals with vocabulary measurements. In addition, the list of 505 expressions mostly comprises the top 2000 words in English, is compatible with other BNC frequency lists, and has a similar purpose to traditional single word lists such as the GSL and AWL, also used in this thesis for vocabulary measurements. I acknowledge the fact that the chosen list was based on the most frequent phrasal expressions in English and this is quite different to the lexical sophistication definition (which is the use of infrequent words, and was used as one of the measures included in my predictive model). I am now using a list based on high frequency. However, the list is based on high frequency expressions used by native speakers. As non-native speakers' use of these expressions tends to be less common, where they are used it can be seen as an indicator of higher proficiency, and learners who use more of them will do so when they are closer to achieving native-like proficiency (higher proficiency).

CHAPTER 5- STUDY 2 METHODOLOGY

5.1 OPERATIONALISATION OF THE TERM COLLOCATIONS

What motivated this second study is the belief that the use of formulaic language (in this instance collocations-phrasal expressions) plays a very important role in lexical knowledge, and the use of collocations is an indication of proficiency (the more proficient a user, the more collocations they understand and use). Therefore, if collocations are measured then we would expect that the more collocations someone has in their essay the higher the grade they attain. If this is the case then another variable can be added to our IELTS ratings' predictability model.

5.2 RESEARCH QUESTIONS

Research Questions

1.   Is there a relationship/correlation between the number of collocations used by the candidates/learners and the ratings?

2.   To what extent can teacher judgements/ratings be predicted (could the model be improved- could the predictive validity increase) by measuring the use of formulaic language (collocations) in the text?

Since Study 2 repeats the procedure followed in Study 1, with modifications including the addition of the two new research questions, the following research questions (that were also investigated in Study 1) are examined (using only the new set of data).

Research questions from Study 1:

•   Do the measures of lexical diversity correlate with the ratings? Which one has the highest correlation?

•   Do the measures of lexical sophistication correlate with the ratings? Which one has the highest correlation?

•   Can we predict the ratings by adding these measures and the formulaic count into a model (based on multiple regression)?

After the completion of Study 1, a request was sent to Cambridge ESOL Examinations in order to gain access to their IELTS Examinations database. For Study 1 the data was collected by myself which was extremely time-consuming and very expensive. Due to the fact that I am self-funded and would need a larger sample of data to improve my model's predictive ability the only way I could get a large amount of IELTS data was directly from the IELTS organisation. Unfortunately, they did not grant me access to their database, so it became necessary for me to search for alternatives.

Therefore, to test this new research question an existing corpus is re-analysed (not the data used for Study 1 but an existing corpus from a previous study in Dubai, which is now publicly available). The data used from Turlik's corpus were the essays (see Section 5.3.2) and the ratings (lexical and holistic) from the IELTS examiners. Below there is an explanation of how the data was collected, prepared and treated by Turlik before been made available to the public (Sections 5.3 and 5.4 refer to how the corpus was collected, how the data was compiled and treated by Turlik). For Study 2 only written data were used, and all steps that were followed for Study 1 are replicated with the exception of the addition of a new variable (to check if this increases the model's predictive ability).

5.3 METHODOLOGY

5.3.1 Participants

This section of the thesis explains how Turlik collected the data (see Turlik 2008: 129-130). The subjects were all female students at Zayed University in the United Arab Emirates. The students on the two year foundation English programme came from state schools where the language of instruction is Arabic, whereas in private schools the language of instruction is English. The subjects socioeconomic status differed, but personal details regarding the parents' occupations, family size or education level could not be collected (even though they would have proven to be useful for the research) due to the fact that some of the subjects were not willing to share such information and the researcher had to accept this, as he claimed that 'it would have been culturally insensitive (as well as against university rules) to press for such information' (Turlik, 2008:130).

It has to be noted that even though the main focus of this research is writing, the students admitted that they are not very familiar with the practice of writing in their schools (Kharma and Hajjaj, 1997:178). It is also worth mentioning that the most the students were ever required to write was a single paragraph, so they were not familiar with the organisation of an essay in English. Writing using the 'Arab style' would be completely different (Turlik, 2008). Turlik explains that there are different conventions learned from students when writing in Arabic. Writing Arabic is different (from right to left) and the structure of the language, grammar and syntax as well as organisational styles are different too (Turlik 2008:123).

In order to enter higher education, students (high school leavers) need to take the CEPA (Common English Proficiency Examination) which includes the CEPA essay used for this study. These essays vary in length and quality depending on the candidate's abilities. Therefore, some essays could consist of just a few sentences or up to 160 words. Furthermore, the fact that Arabic is written from right to left does seem to influence the students' spelling (they often misspell words). This issue is addressed in the thesis after the data treatment section.

5.3.2 Essays

The data of this corpus were collected (by Turlik) over a period of two years at Zayed University, Dubai. The 340 essays collected over that period were written by 42 students (there were originally 44 students but two of them, student 6 and 9 had to be excluded –see below for details or refer to Turlik, 2008: 130). The first writing text used was the CEPA (Common English Proficiency Assessment) which students take on completion of high school, and the final essay was the last/final essay written by the students once they had completed their foundation (Readiness) English course at Zayed University.

Over the two year data collection period there were almost no changes to the examination format, and an attempt was made to control the numbers of independent variables that could affect the results. The issue of genre was important, as having various genre samples could have an effect on the results.

The first essay collected for the 44 subjects was the Common English Proficiency Exam (CEPA). This exam lasts 2¼ hours and needs to be taken by students that wish to enter university. It tests students reading comprehension, grammar and composition, and its integrity is not questioned.

The period that the researcher could have reliable data was two years which was the period of time students spent on the Readiness Programme. Whether the participant/student passed or failed was not important, but it was essential for a student to complete Level 8 (if they were to be to be included in the sample). Level 2 was the lowest level students had to achieve to be accepted, and if someone did not complete Level 8 their work was excluded from the study. All the essays were written under examination conditions and they were all responses to examination questions.

According to Turlik (2008), there are some gaps in the data set because some essays were not found in students' exam packs. Even though this was problematic, students were not rejected because of the omission of one of their essays. However, two students (Student 7 and 9) were rejected due to the fact that their Level 8 essays were found to have been destroyed. Turlik also explains how the genre variable had been accommodated:

> 'Genre, arguably a prominent variable, has been accommodated insofar as there is a limited number of titles and, over the time the students spent in Readiness, most answered at least one from each of the title groups and if the essays are grouped under four, more general genre headings, then a slightly different picture emerges.' (Turlik, 2008:131/132)

There was of course an expected rate of attrition from students that left the university for a plethora of reasons.

5.3.3 Corpus

After the procedure of data collection, the corpus consisted of 340 essays (including CEPA) from 42 students collected by Turlik during the period between March 2003 and September 2005.

Out of the 340 essays, 41 were CEPA essays and the rest (299) were end-of-course formal assessments given at the end of a ten week 'term', all written under examination conditions (Turlik, 2008).

It has to be noted that even though this corpus has been used in a previous study, it is now available for the public (see link in previous chapter), and my study will be a complete re-analysis of the data collected by Turlik (2008).

5.3.4 Raters' measures

In this section an explanation is provided of how the ratings were made (see Turlik 2008: 135). In order to be able to compare human ratings with mathematically derived measures of lexical richness, inter-rater consistency should be checked based on established and proven criteria. It was decided that the best criteria to use would be the actual IELTS band descriptors. Turlik was granted permission from UCLES/Cambridge for trained IELTS raters (at Zayed University) to provide two IELTS ratings for each essay: a rating for lexical range and accuracy and a holistic rating using the IELTS band descriptors for Writing (see Appendix 2).

The choice for this specific band descriptor is justifiable as the model that I am trying to create will be predicting IELTS scores. Most of the variables that the model will consist of match the criteria from the IELTS Writing band descriptor for lexical range and accuracy. In order to ensure anonymity the 340 essays were given to the two raters without the authors' names (only reference numbers) and there was no indication of the sequence in which they were written. The two raters were given instructions on how to rate the essays. They were asked to provide two scores/ratings: an initial holistic rating and then a score for lexical range and accuracy. They were asked to rate the essays as if they had been submitted by IELTS candidates.

5.4 TREATMENT OF DATA

Turlik (2008) ensured that all the essays were transcribed by a specialist agency into computer readable texts to be analysed in CLAN and other programmes. The procedures that were followed (for consistency purposes) for the editing of the essays can be seen in Appendix 9.

5.4.1 Handwriting and Spelling

Handwriting and quality of presentation can affect raters' judgement (G Yu 2009). This is the main reason that all essays were typed before they were given to raters. Spelling mistakes were corrected in an effort to limit the percentage of low frequency words. If spelling mistakes had not been corrected, then during the various calculations they would have been counted as infrequent (more sophisticated) words, even if they had just been high-frequency words (just spelled wrongly). The procedure followed is quite clear, but according to Turlik (2008:133), 'spelling was corrected, as the view was taken that spelling is part of the learning process and a word used in the correct context but spelt incorrectly should be acceptable'. In addition, according to Schmitt, spelling does not seem to be an issue for advanced-proficient users (Schmitt, 1998a).

5.5 RE-ANALYSIS OF THE DATA

In this section, an explanation of the chosen measures that were included in my own analysis of the data is provided, and certain procedures and calculations that were part of my own methodology and research design are explained.

5.5.1. Lemmatisation of data

The data was not lemmatised for the same reasons explained in Study 1 (see Chapter 3 for a detailed explanation of the non-lemmatisation decision).

5.5.2 Measures

The measures used to calculate the essays' lexical richness were the following:

Lexical Sophistication Measures	Lexical Diversity Measures	Raters' Judgements
Number of types	Number of tokens	Lexical range and Accuracy
Guiraud Advanced	Guiraud	Holistic Rating
P_Lex (Lambda values)	TTR (Type-Token Ratio)	
Formulaic count (Phrasal expressions count)	Malvern and Richards D	

The measures included were the same as Study 1 (please refer back to Chapter 3 for discussion of selection) with the addition of a new depth of knowledge measure (Formulaic count) to help improve the predictive validity of my model.

5.5.3 Equipment and software

All ratings were put in an SPSS file along with the scores of lexical richness of both quantitative and qualitative measures. The SPSS file consists of the following variables:

Essay number

Number of types

Number of tokens

Guiraud

Guiraud advanced

'D' value

TTR

P_Lex

Phrasal expressions

Lexical range and accuracy ratings

Holistic ratings

5.5.4 Calculations

All the variables were entered in an SPSS file. The variables were: the student number, the holistic and lexical ratings from the raters for each essay and the averages, all the calculations from the lexical richness measures (lexical diversity and sophistication), and the formulaic count. All calculations for all the measures of lexical richness were carried out using the same methods as Study 1 (Pilot study). Formulaic count was carried out manually (using the PHRASE List).

The vocd command was used in CLAN to calculate the number of types, tokens, TTR and D values of the essays. Words that were repeated or place names were excluded from the count. There were 44 students initially, but after the exclusion of Students 7 and 9 there were 42 left. The first essay is missing from the data of Student 3. In addition, the seventh essay is missing from the data of 3 students: Students 7, 10 and

23. Essay 38.1 was also excluded because it was not rated (too short). Some students' essays were too short, and therefore the calculations were not possible (TTR and D could not be calculated in CLAN). This is the list of essays: 5.1, 13.1, 16.1, 20.2, 25.1, 42.2 and 43.1.

Guiraud Advanced was calculated by using Eugene Mollet's programme (Personal communication with Daller), which is based on two wordlists (please refer back to Chapter 3- section 3.3.4- for a description of the wordlists and short discussion).

P_Lex was used to calculate lambda values. A short description of how the programme works is provided in section 3.2.3 in Chapter 3. You can find the list of what were considered sophisticated words in Appendix 10. This task was not as straightforward as it seems, as there were some words that were harder to categorise. According to the instructions in the P_Lex manual, easy words should be words that belong in the Level 1 list and their basic derivatives. This basic derivatives definition was not clear. For example, I had problems deciding whether to consider the word *preparatory* as an easy or hard word. The word *prepare* belongs in the list but my own intuition and experience as an English teacher led me to consider the word *preparatory* as a hard word (even though it is a derivative of the word *prepare*, which belongs in the Level 1 list). Students would commonly know and use the word *prepare,* but not many would be able to produce and use the word preparatory. This was one of the main reasons for my decision to count the derivatives of the words of the lists as different types (not lemmatised data).

Instances of formulaic language were operationalised by Martinez and Schmitt's list (2012) of the 505 most frequent multiword expressions in English, and were counted manually after contacting both the authors and discovering that up to that moment there is was automatic method of calculation.

I expected all indices to increase with level and scores. The same applied to formulaic count. However, while calculating the lambda values, I noticed that sometimes a student with low formulaic count would get a higher lambda value than someone with a high formulaic count (that got a lower lambda value). After further analysis (looking at which phrases were used not just how many) I found that students with lower lambda

values and a high formulaic count would usually only use a couple of phrases many times, while others with a lower count would use a variety of different phrases. This explained the fact that higher lambda scores were obtained by students that received higher scores (were put on a higher level/band) and was the opposite of what was expected and hypothesised.

CHAPTER 6- PRESENTATION OF RESULTS & DISCUSSION

In this Chapter the results of Study 2 are presented and discussed.

6.1 DESCRIPTIVE STATISTICS

6.1.1 Treatment of Data

First of all, the data from the students that were excluded (student 6 & 9) was omitted from the analysis. Therefore, the data used was from the 42 remaining students.
The following calculations were computed using SPSS for all the essays:

6.1.2 Inter-rater reliability

The first thing we want to investigate is inter-rater reliability and this can be investigated by calculating Cronbach's Alpha, which is a coefficient of internal consistency. First, the reliability of the holistic ratings was checked. The results can be found in Table 6.1.

Table 6.1: Cronbach's Alpha for Holistic Ratings

Reliability Statistics

Cronbach's Alpha	N of Items
.667	2

The value of 0.667 is questionable. Based on the value of 0.667, the internal consistency of the two raters is questionable. As a rule of thumb, the value needs to be over 0.7 for the results to be considered acceptable to work with (Kline, 1999; George and Mallery, 2003).

Table 6.2: Cronbach's Alpha Values

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.8 \leq \alpha < 0.9$	Good

$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

(Kline, 1999; George and Mallery, 2003)

This could be a limitation of the study. The low alpha value indicates that there is a difference between the raters. The value could be low because, according to Revelle and Zinbarg (2009), 'Alpha is not robust against missing data', and there was some data missing from my sample.

Cronbach'alpha was also calculated for the lexical ratings (see Table 6.3).

Table 6.3: Cronbach's Alpha for Lexical Ratings

Reliability Statistics

Cronbach's Alpha	N of Items
.717	2

The value of 0.717 is acceptable so we can continue with the research.

In order to further investigate the reliability between the two raters, a paired t-test was chosen (see Table 6.4).

Table 6.4: Comparison between Ratings (paired samples t-test)

Paired Samples Test

		Paired Differences			t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	Rater1H - Rater2H	.314	.877	.052	6.03	282	p < 0.001
Pair 2	Rater1L - Rater2L	.760	.922	.055	13.86	282	p < 0.001

Both in terms of the lexical and in terms of the holistic ratings, the mean difference rating between Rater 1 and Rater 2 is significantly different from zero, and positive at

the 0.1% significance level (in fact p value is less than .0005), in both cases therefore, Rater 1 seems to give higher rates on average than Rater 2.

6.1.3 The Means of the Raters' Lexical and Holistic Ratings

As can been seen from Table 6.5, the total number of essays examined, after all the exclusions, is 283 (please see Section 5.4- under calculations you can find details on how only 283 essays remained). The means of Rater 1 are higher than Rater 2 and the standard deviation is lower, which indicates that the homogeneity of Rater 1 is higher than that of Rater 2.

Table 6.5: Descriptive Statistics for Ratings

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Rater1H	283	2	8	4.13	0.71
Rater2H	283	1	8	3.82	1.02
Rater1L	283	2	8	4.62	0.80
Rater2L	283	1	8	3.86	1.13

6.1.4 Correlations

Correlations between Measures of Lexical Diversity

A scatterplot was first constructed to visually check the correlations between the measures of lexical diversity (see Appendix 11). There are correlations between all the measures as expected but two pairs of measures stand out: one is regarding the relationship of TTR and tokens which seems to have a strong negative correlation (the more tokens the lower the TTR values). This is not surprising, as the TTR is directly affected by the number of tokens since it is calculated as the number of types divided by the number of tokens (Types / Tokens). The other point worth noting is the strong positive correlation between the measures D and Guiraud, suggesting that essays with high values are more likely to secure high values of the Guiraud measure.
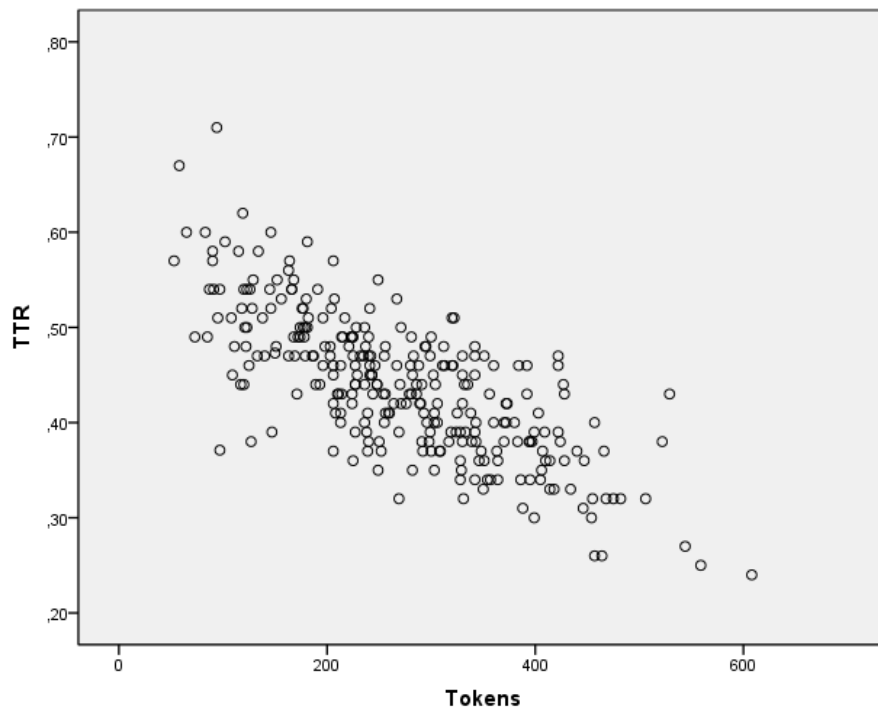
Figure 6.1. Correlation between TTR and Number of Tokens (Scatterplot)
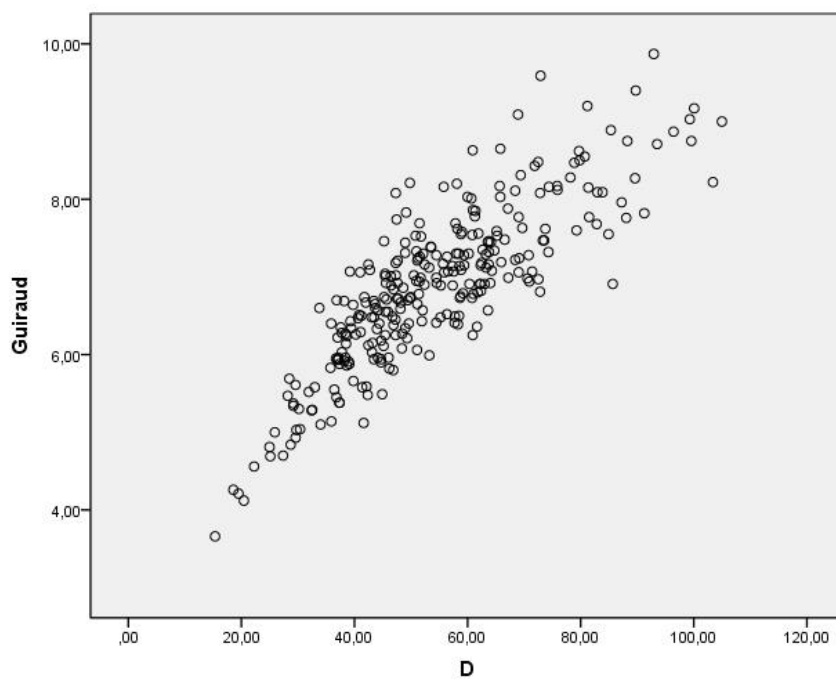
Figure 6.2. Correlation between D and Guiraud (Scatterplot)

The pair wise correlations can be seen in the following table:

Table 6.6: Correlations between Measures of Lexical Diversity

Correlations		Tokens	D	TTR	Guiraud
Tokens	Pearson Correlation		.451**	-.747**	.553**
	Sig. (2-tailed)		.000	.000	.000
D	Pearson Correlation			.095	.843**
	Sig. (2-tailed)			.110	.000
TTR	Pearson Correlation				.041
	Sig. (2-tailed)				.494
Guiraud	Pearson Correlation				
	Sig. (2-tailed)				
**. Correlation is significant at the 0.01 level (2-tailed).					

The highest correlation is Guiraud and D (0.84) which seem to increase at the same level (when one's values increase, the other's values seem to increase also). There is also a high negative correlation between TTR and tokens (-0.74), which shows that when one increases (tokens) the other one's values are expected to drop (as was also observed in Figure 6.1).

Correlations between Measures of Lexical Sophistication

The measures of lexical sophistication were examined through a series of scatterplots (see Appendix 12). While all correlation coefficients between the measures of lexical sophistication were significantly different from zero, they were not generally as strong as the correlations exhibited by the lexical diversity measures, as can be seen in the following table.

Table 6.7: Correlations between Measures of Lexical Sophistication

Correlations		Types	P_Lex	GuiAdv	Form
Types	Pearson Correlation		.235**	.441**	.435**
	Sig. (2-tailed)		.000	.000	.000
P_Lex	Pearson Correlation			.581**	.164**
	Sig. (2-tailed)			.000	.006
GuiAdv	Pearson Correlation				.098
	Sig. (2-tailed)				.102
Form	Pearson Correlation				
	Sig. (2-tailed)				

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

Table 6.8: Correlations between Measures of Lexical Richness and Lexical Ratings

Correlations		Rater 1L	Rater 2L	Tokens	Types	D	TTR	Guiraud	P_Lex	GuiAdv	Form	Raters AvL
Rater1L	Pearson Correlation		.593**	.602**	.601**	.332**	-.441**	.423**	.352**	.285**	.355**	.850**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Rater2L	Pearson Correlation			.608**	.621**	.353**	-.368**	.488**	.301**	.413**	.331**	.928**
	Sig. (2-tailed)			.000	.000	.000	.000	.000	.000	.000	.000	.000
Tokens	Pearson Correlation				.897**	.451**	-.747**	.553**	.149*	.247**	.485**	.678**
	Sig. (2-tailed)				.000	.000	.000	.000	.012	.000	.000	.000
Types	Pearson Correlation					.712**	-.444**	.855**	.235**	.441**	.435**	.686**
	Sig. (2-tailed)					.000	.000	.000	.000	.000	.000	.000
D	Pearson Correlation						.095	.843**	.202**	.345**	.255**	.386**
	Sig. (2-tailed)						.110	.000	.001	.000	.000	.000
TTR	Pearson Correlation							.041	-.024	.098	-.406**	-.446**
	Sig. (2-tailed)							.494	.691	.099	.000	.000
Guiraud	Pearson Correlation								.281**	.564**	.242**	.517**
	Sig. (2-tailed)								.000	.000	.000	.000
P_Lex	Pearson Correlation									.581**	.164**	.360**
	Sig. (2-tailed)									.000	.006	.000
GuiAdv	Pearson Correlation										.098	.402**
	Sig. (2-tailed)										.102	.000
Form	Pearson Correlation											.381**
	Sig. (2-tailed)											.000
RatersAvL	Pearson Correlation											
	Sig. (2-tailed)											

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 6.9: Correlations between Measures of Lexical Richness and Holistic Ratings

Correlations		Tokens	Types	D	TTR	Guiraud	P_Lex	GuiAdv	Form	Rater 1H	Rater 2H	Raters AvH
Tokens	Pearson Correlation		.897**	.451**	-.747**	.553**	.149*	.247**	.485**	.584**	.579**	.663**
	Sig. (2-tailed)		.000	.000	.000	.000	.012	.000	.000	.000	.000	.000
Types	Pearson Correlation			.712**	-.444**	.855**	.235**	.441**	.435**	.602**	.585**	.675**
	Sig. (2-tailed)			.000	.000	.000	.000	.000	.000	.000	.000	.000
D	Pearson Correlation				.095	.843**	.202**	.345**	.255**	.399**	.311**	.395**
	Sig. (2-tailed)				.110	.000	.001	.000	.000	.000	.000	.000
TTR	Pearson Correlation					.041	-.024	.098	-.406**	-.369**	-.353**	-.410**
	Sig. (2-tailed)					.494	.691	.099	.000	.000	.000	.000
Guiraud	Pearson Correlation						.281**	.564**	.242**	.465**	.448**	.519**
	Sig. (2-tailed)						.000	.000	.000	.000	.000	.000
P_Lex	Pearson Correlation							.581**	.164**	.302**	.271**	.322**
	Sig. (2-tailed)							.000	.006	.000	.000	.000
GuiAdv	Pearson Correlation								.098	.334**	.378**	.409**
	Sig. (2-tailed)								.102	.000	.000	.000
Form	Pearson Correlation									.260**	.286**	.313**
	Sig. (2-tailed)									.000	.000	.000
Rater1H	Pearson Correlation										.536**	.823**
	Sig. (2-tailed)										.000	.000
Rater2H	Pearson Correlation											.920**
	Sig. (2-tailed)											.000
RatersAvH	Pearson Correlation											
	Sig. (2-tailed)											

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

The same variables/measures (Tokens and Types) seem to have the highest correlations with the holistic ratings.

Correlations between Formulaic Count and Ratings (Lexical and Holistic)

The correlation between the formulaic count and the ratings - both lexical and holistic- are quite low (see Tables 6.8 and 6.9).

6.2 REGRESSION ANALYSES AND INFERENCE

6.2.1 Predictive Model for Lexical Ratings

After all the correlations were checked the whole population was analysed using a multiple regression analysis using all the previous variables (lexical diversity, sophistication and formulaic count) as predictor variables for the IELTS ratings. It is stressed that since all the measures of lexical richness were used as predictors in the

regression model for the lexical ratings, variance inflation factors were calculated to check for the presence of multicollinearity:

Table 6.10: Regression Coefficients from full model (Lexical ratings)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.775	.782		3.547	.000		
	Tokens	.002	.002	.191	.727	.468	.024	42.136
	Types	.010	.011	.392	.917	.360	.009	110.941
	D	-.004	.004	-.075	-.907	.365	.239	4.185
	TTR	-1.284	1.316	-.112	-.976	.330	.124	8.050
	Guiraud	.032	.216	.038	.146	.884	.024	41.505
	P_Lex	.377	.101	.191	3.712	.000	.618	1.618
	GuiAdv	.225	.181	.077	1.240	.216	.423	2.362
	Form	.008	.011	.035	.731	.465	.716	1.396

a. Dependent Variable: RatersAvL

As can be seen in Table 6.10, there seems to be a problem with the variable Types (VIF was too high). Therefore, another regression was carried out which excluded the variable Types.

Table 6.11: Regression Coefficients with variable 'Types' removed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.291	.577		3.969	.000		
	Tokens	.003	.001	.400	3.025	.003	.094	10.653
	D	-.004	.004	-.073	-.881	.379	.239	4.181
	TTR	-1.565	1.279	-.137	-1.223	.222	.131	7.614
	Guiraud	.212	.091	.255	2.326	.021	.136	7.347
	P_Lex	.371	.101	.189	3.665	.000	.620	1.612
	GuiAdv	.229	.181	.079	1.263	.208	.424	2.360
	Form	.010	.011	.041	.875	.383	.732	1.367

a. Dependent Variable: RatersAvL

The Tokens' Variance Inflation Factor (VIF) value is above 10. To prevent collinearity issues the value needs to be under 10. The VIF value of Tokens is 10.653, so it needs to be excluded. As a rule of thumb, variables higher than 10 need to be excluded from the model.

A third regression followed without the inclusion of the variable Tokens:

Table 6.12: Regression Coefficients with variable 'Tokens' removed

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	3.503	.422		8.307	.000		
	D	-.001	.004	-.024	-.295	.769	.249	4.023
	TTR	-5.091	.535	-.445	-9.515	.000	.773	1.293
	Guiraud	.365	.077	.440	4.745	.000	.197	5.084
	P_Lex	.325	.102	.165	3.195	.002	.635	1.575
	GuiAdv	.289	.183	.099	1.581	.115	.429	2.332
	Form	.013	.011	.055	1.153	.250	.739	1.354

a. Dependent Variable: RatersAvL

As can now be seen in Table 6.12, all the VIF values are lower than 10. However, now another problem arises. The variable D needs to be excluded from the model because it is not statistically significant (P value is high p=0.769).

Therefore, another regression follows with the exclusion of D as a predictor of lexical ratings:

Table 6.13: Regression Coefficients with variable 'D' removed

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	3.575	.341		10.485	.000		
	TTR	-5.124	.522	-.448	-9.809	.000	.809	1.237
	Guiraud	.346	.043	.417	8.091	.000	.634	1.578
	P_Lex	.321	.101	.163	3.187	.002	.643	1.556
	GuiAdv	.306	.173	.105	1.768	.078	.477	2.098
	Form	.012	.011	.053	1.123	.262	.756	1.323

a. Dependent Variable: RatersAvL

There is now another variable that is not statistically significant and needs to be excluded, the Formulaic count (p=0.262). The results of the next regression analysis can be found in the following table.

Table 6.14: Regression Coefficients with variable 'Form' (Formulaic Count) removed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.665	.332		11.052	.000		
	TTR	-5.369	.475	-.469	-11.310	.000	.980	1.021
	Guiraud	.358	.041	.432	8.663	.000	.678	1.475
	P_Lex	.335	.100	.170	3.351	.001	.653	1.532
	GuiAdv	.290	.173	.100	1.679	.094	.480	2.084

a. Dependent Variable: RatersAvL

Guiraud Advanced (GuiAdv) seems to have a high p value (p=0.094) therefore it was excluded from the model in order to improve its validity. The results from the last regression can be seen below.

Table 6.15: Regression Coefficients with variable 'Guiraud Advanced' removed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.461	.309		11.183	.000		
	TTR	-5.265	.472	-.460	-11.150	.000	.997	1.003
	Guiraud	.394	.036	.475	11.049	.000	.919	1.088
	P_Lex	.426	.085	.216	5.035	.000	.920	1.087

a. Dependent Variable: RatersAvL

All the remaining factors/variables (TTR, Guiraud and P_Lex) are highly significant in predicting the lexical ratings. It also seems that even though the TTR and Guiraud are both measures of lexical diversity, they seem to explain different aspects of the ratings because their VIF values are very low.

Therefore TTR, Guiraud and P_Lex (two measures of lexical diversity and one of lexical sophistication) are the three measures that can explain 52. 8% of the variability in the lexical ratings.

Table 6.16: Final Regression Model (Lexical Ratings) Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.727[a]	.528	.523	.58851

a. Predictors: (Constant), P_Lex, TTR, Guiraud

The fitted regression model for the lexical ratings is as follows:

IELTS Lexical rating: $3.461 - 5.265*TTR + 0.394*Guiraud + 0.426*P\_Lex$

The standardised residuals seem to satisfy the necessary regression assumptions including normality. The majority of standardised residuals fall between -2 and +2, while the normal curve closely fits their histogram (see Figure 6.3 below).

Visualising the model fit and the regression line is not possible given that there are three explanatory variables (TTR, Guiraud, P_Lex) that predict the lexical rating (dependent variable).
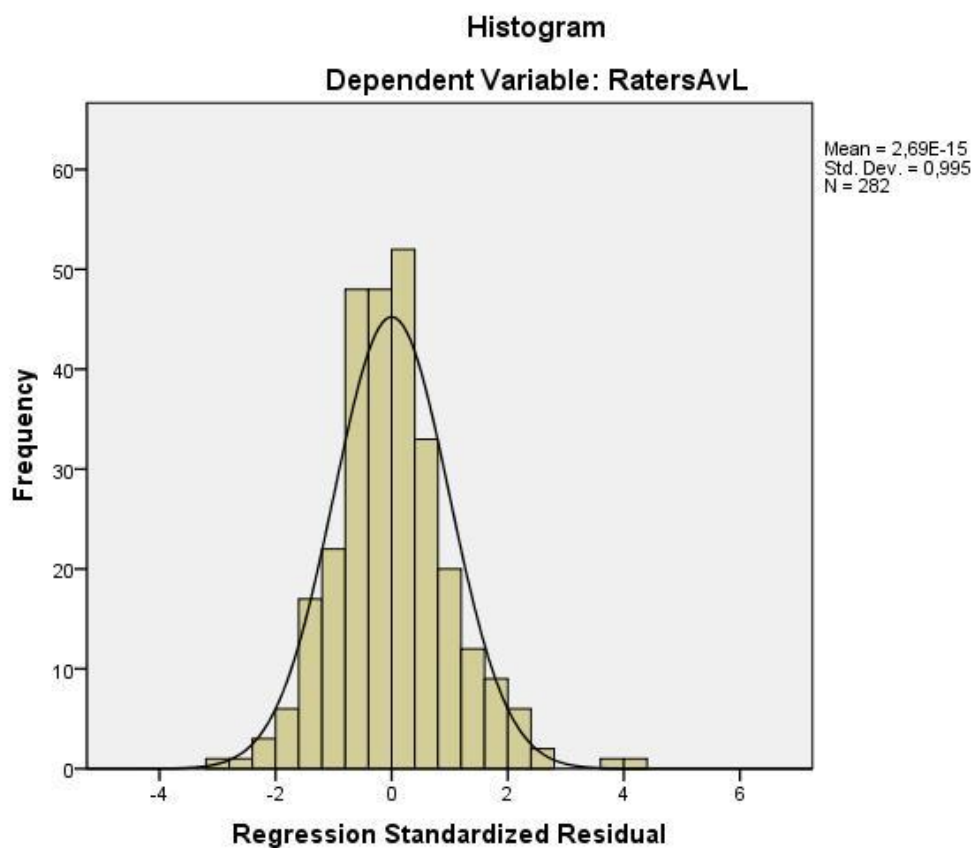
Figure 6.3: Histogram of Standardised Residual from Lexical Ratings final model

6.2.2 Predictive Model for Holistic Ratings

The same steps were followed for finding the model for predicting the holistic ratings. A regression analysis using backward elimination was carried out using all the measures as predictors of holistic ratings.

Table 6.17: Regression Coefficients from full model (Holistic ratings)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	2.370	.711		3.333	.001		
	Tokens	.003	.002	.402	1.476	.141	.024	42.136
	Types	.004	.010	.201	.455	.649	.009	110.941
	D	-.001	.004	-.032	-.370	.712	.239	4.185
	TTR	-.430	1.196	-.043	-.360	.719	.124	8.050
	Guiraud	.039	.197	.053	.198	.843	.024	41.505
	P_Lex	.242	.092	.140	2.626	.009	.618	1.618
	GuiAdv	.311	.165	.122	1.889	.060	.423	2.362
	Form	-.008	.010	-.037	-.746	.457	.716	1.396

a. Dependent Variable: RatersAvH

For the same reasons explained above (high VIF values), the variables Types and then Tokens were removed in order to improve the model. Then, another regression analysis was calculated:

Table 6.18: Regression Coefficients with variables 'Types' and 'Tokens' removed

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	3.505	.386		9.084	.000		
	D	.001	.004	.031	.365	.715	.249	4.023
	TTR	-4.496	.490	-.448	-9.183	.000	.773	1.293
	Guiraud	.291	.070	.400	4.136	.000	.197	5.084
	P_Lex	.188	.093	.109	2.018	.045	.635	1.575
	GuiAdv	.380	.167	.149	2.273	.024	.429	2.332
	Form	-.003	.010	-.016	-.324	.746	.739	1.354

a. Dependent Variable: RatersAvH

The variables Form (Formulaic count) and D need to be excluded (not statistically significant coefficients).

Table 6.19: Regression Coefficients with variables 'Form' and 'D' removed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.403	.303		11.235	.000		
	TTR	-4.404	.434	-.439	-10.158	.000	.980	1.021
	Guiraud	.309	.038	.426	8.197	.000	.678	1.475
	P_Lex	.188	.091	.109	2.059	.040	.653	1.532
	GuiAdv	.365	.158	.143	2.313	.021	.480	2.084

a. Dependent Variable: RatersAvH

In Table 6.19 we can see that P_Lex has a p value= 0.040. If we were using a .05 significance level for the model, all four remaining variables (TTR, Guiraud, P_Lex and GuiraudAdvanced) could be used in predicting the holistic ratings. However, if we reduce the significance level to .01 we get different results:

Table 6.20: Regression Coefficients with variable 'P_Lex' removed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.553	.296		12.014	.000		
	TTR	-4.495	.434	-.448	-10.360	.000	.990	1.010
	Guiraud	.304	.038	.418	8.023	.000	.682	1.467
	GuiAdv	.540	.134	.211	4.039	.000	.676	1.479

a. Dependent Variable: RatersAvH

Now, after the exclusion of P_Lex, all the coefficients' significance values are lower than 1% ($p < 0.01$) indicating very strong level of significance (even lower than .05).

These three variables seem to explain 48.6 % of the variability in the holistic ratings.

Table 6.21: Final Regression Model (Holistic Ratings) Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.697[a]	.486	.480	.53880

a. Predictors: (Constant), GuiAdv, TTR, Guiraud

The fitted regression model for predicting the holistic ratings is the following:

IELTS Holistic Rating= 3.553-4.495*TTR +0.304*Guiraud+ 0.540*Guiraud Advanced.

Checking the results in a histogram reveals an unusual detail.
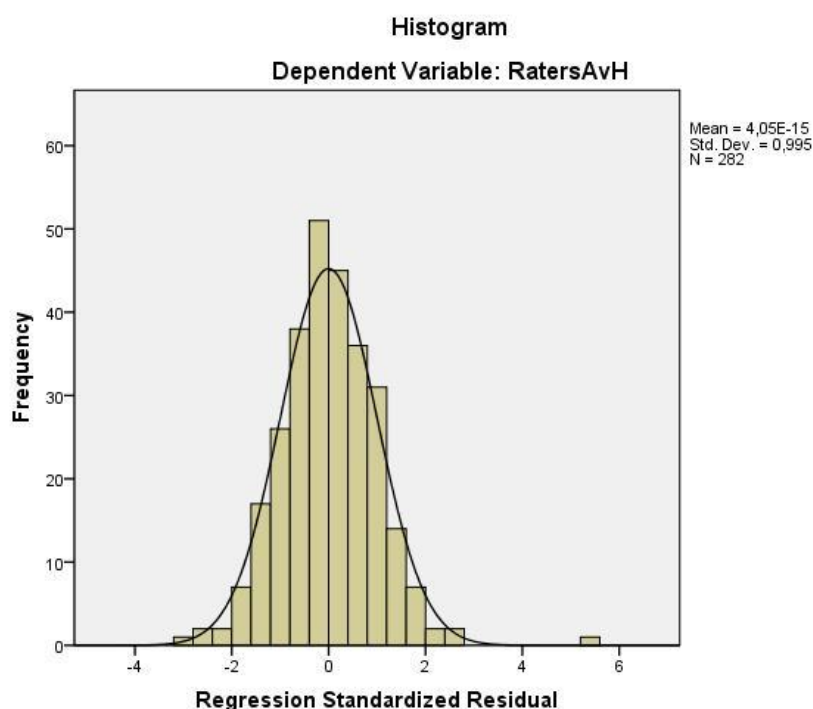
Figure 6.4: Histogram of Standardised Residual from Holistic Ratings final model

There seems to be a student's essay (outlier) that received higher marks than the model predicted (the model is underestimating the value). I revisited the data and checked this particular essay; it seems that it could be considered as an extreme case. This student was extremely good and received very high marks. In addition, all of her lexical richness values were higher than anyone else's. This can be considered an influential point and the model could be improved if this case was excluded (see Table 6.22 and Figure 6.5 below). These are the results after the exclusion of that case:

Table 6.22: Final Regression Model (Holistic Ratings) Summary with outlier removed

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.702[a]	.492	.487	.50790

a. Predictors: (Constant), GuiAdv, TTR, Guiraud

The model was slightly improved. Both the R^2 and the adjusted R^2 measures of goodness-of-fit have increased and the model can now explain 49.2% of the variability in the holistic ratings. In addition, the residuals now seem to satisfy the necessary regression assumptions, including normality, with their histogram following closely the normal curve (see Figure 6.5).
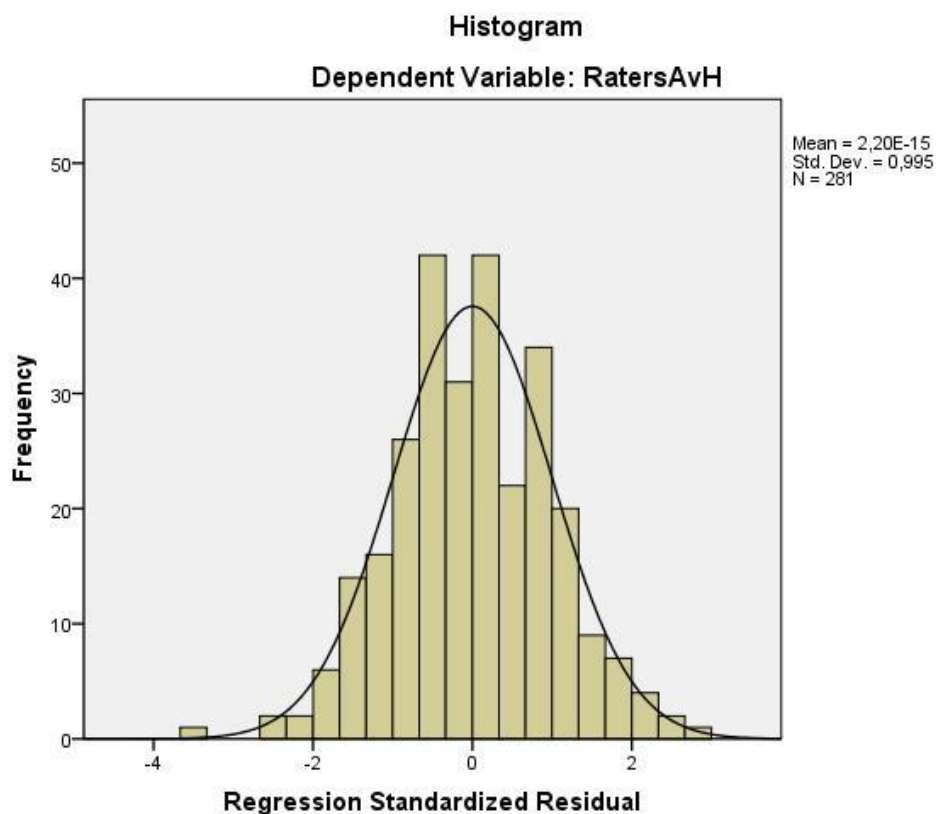
Figure 6.5: Histogram of Standardised Residual from Holistic Ratings final model with outlier removed

6.3 DISCUSSION OF RESULTS

In this section the results from Study 2 are discussed. Each research question is addressed separately, presenting the results and discussing any implications and relations with literature and expected findings.

To start with, McCarthy and Jarvis (2010) argue about what is called 'convergent validity' of the measures which means that measures of theoretically similar constructs should be highly intercorrelated. They believe that there should be correlations between measures of the same aspect (between measures of lexical diversity and between measures of lexical sophistication). Therefore, this is what I expect to find in my research.

Firstly, all the correlations between the measures of lexical diversity and lexical sophistication were significant. However, the correlations between the measures of lexical sophistication were not as strong as the correlations shown by the lexical diversity measures.

The main correlation (between the measures of lexical diversity) that stands out from the results is that regarding the relationship of TTR and tokens, which seems to have a strong negative correlation (the more tokens the lower the TTR values). This was not at all surprising, as the TTR is directly affected by the number of tokens since it is calculated as the number of types divided by the number of tokens (Types / Tokens). This was supported by the literature (Malvern and Richards, 1997; Malvern et al, 2004). TTR decreases with increasing tokens as speakers repeat themselves. TTR was not one of the variables that I expected (due to supporting literature by Mayor et al, 2002 who found no apparent relationship between IELTS band scores and TTR) to correlate highly with the ratings and be one of the predictors used in the model . Another point worth noting is the strong positive correlation between the measures D and Guiraud (0.84) that suggest that essays with high values are more likely to yield high values of the Guiraud measure.

Do the measures of lexical diversity correlate with the ratings? Which one has the highest correlation?

All the measures of lexical diversity do correlate with the ratings. The measure of lexical diversity that has the highest correlation with the lexical ratings of both raters is the number of tokens (for Rater 1: $r =. 602$ and Rater 2: $r =. 608$). The same appears to apply for the holistic ratings.

Do the measures of lexical sophistication correlate with the ratings? Which one has the highest correlation?

All the measures of lexical sophistication correlate with the ratings. The measure of lexical sophistication that has the highest correlations with the lexical ratings of both raters is the number of types (Rater 1: $r =. 601$ and Rater 2: $r =. 621$). The same rule appears to apply for the holistic ratings.

The results from the correlations between the measures of lexical richness and the teacher ratings are supported in previous literature by Banerjee et al. (2004). Their study showed a correlation between the number of tokens and the number of types and the IELTS overall scores. The higher mark the candidate achieved in the IELTS exam, the more tokens and types were found in their speech. This is also supported by Mayor et al. (2002), who argued that word count (tokens) is one of the strongest predictors of band score in IELTS Writing task 2 performance. Furthermore, Hsu (2007) found that the word count (total words) in the essays correlated significantly with writing scores. Cobb's (2002) study also showed that the more words (tokens) produced by a learner the higher the level they achieved.

Is there a relationship/correlation between the number of collocations used by the candidates/learners and the ratings?

There is a significant correlation between the number of collocations (formulaic count) used by the candidates and the ratings. However, this is not the high correlation that was expected. This finding contrasts with the results of Read and Nation's (2002) study, which after conducting a more qualitative analysis, found that higher band candidates

used more formulaic language in their speech. However, one of the differences between this study and Read and Nation's study is the fact that they focused on oral speech (oral data) whereas my study (Study 2) focuses on written data. Oral language is formulaic by nature so this could be a possible explanation regarding their results (see discussion in Chapter 4, Section 4.2.1 on how formulaic language is more commonly found in oral than written speech). In addition, Read and Nation explain that the examiners have guidelines on how to form questions for the Speaking test and it is quite formulaic in nature as well. Another possible explanation is the operationalisation of formulaic language. Formulaic language is difficult to define and even though both studies used the definition by Wray (2002; 2008) for what formulaic language is, it was operationalised in different ways. Read and Nation (2002) chose formulaic language intuitively (which makes their approach subjective in nature). They selected transcripts and marked words, phrases or longer sequences of vocabulary that they regarded as formulaic (Read and Nation, 2002). Therefore, they proposed that due to the exploratory nature of their research their findings should be regarded as suggestive and not conclusive. On the other hand, I used Martinez and Schmitt's List which was not available at the time Read and Nation conducted their study. However, this list could be more appropriate for oral data, so it may have shown different results if used on a different corpus (of oral data) and could be something to explore in the future (see below for more on this issue).

<u>To what extent can teacher judgements/ratings be predicted (could the model be improved- could the predictive validity increase) by measuring the use of formulaic language (collocations) in the text?</u>

The formulaic count variable was added to the model after suggestions found in existing literature. For example, Bogaards (2000) stated that even though the size of vocabulary is important, other aspects such as collocations are equally important (Bogaards 2000). In addition, the use of collocations is a linguistic characteristic of academic texts according to Biber, Conrad, Reppen, Byrd and Helt (2002). It seems that, even though it was expected that formulaic count would improve my model's predictive validity, the model was not improved by the formulaic count variable. The formulaic count had to be excluded from the model (non-statistically significant coefficient). The lexical ratings model consists of three variables: two of lexical diversity and one of lexical

sophistication. TTR, Guiraud and P_Lex are the three measures (variables) that can explain 52. 8% of the variability in the lexical ratings as shown in the following model (see also page 151):

IELTS Lexical rating: 3.461 – 5.265*TTR + 0.394*Guiraud + 0.426*P_Lex

In addition, holistic ratings can be predicted by the same two lexical diversity measures (TTR and Guiraud), but using a different measure of lexical sophistication, Guiraud Advanced. A model consisting of these three measures can explain 49.2% of the variability in the holistic ratings, as shown in the model below:

IELTS Holistic Rating= 3.553-4.495*TTR +0.304*Guiraud+ 0.540*Guiraud Advanced.

Two of the strongest predictors of IELTS Writing ratings are measures of lexical diversity. Read and Nation (2002) found the opposite when investigating whether IELTS Speaking ratings could be predicted by measures of lexical diversity. It was suggested that measures of lexical diversity could not predict proficiency levels. Guiraud has been very much criticised in the past, however, it seems to be performing well in this data. This is quite a surprising result, but is in line with most other recent studies (Michael Daller, personal communication). Guiraud, according to Van Hout and Vermeer (2007), correlates better with language ability than other measures (such as D) and this was confirmed in my study too. This finding was also in line with Fatahipour's study (2012) who reported that Guiraud proved to be a better measure of language ability since there was a higher correlation of Guiraud and language ability than D and language ability.

Therefore, we can predict IELTS writing ratings with the use of lexical richness measures, but not with the addition of the formulaic count variable. It could thus be suggested that adding a measure of vocabulary depth of knowledge did not improve the model at this stage (collocations are considered to be an aspect of depth of knowledge- Beglar and Hunt, 1999). A possible explanation as to why the formulaic sequences variable did not improve my model at this stage could be the fact that there are more formulaic sequences, and collocations in particular, found in spoken language because

as Simpson-Vlach and Ellis (2010:488) explain: 'Speech is constructed in real time and this imposes greater working memory demands than writing, hence the greater the need to rely on formulas'.

It is worth noting that D (which I expected to be one of the predictors for ratings) was not included in the model. This is quite a disappointing and unexpected result. According to recent research (Crossley et al. 2011a), D seemed to be a good predictor of teacher ratings, so it is surprising that it had to be excluded from this study model. In Crossley et al.'s study, D seemed to explain 34% of the variance in human judgements. One possible explanation as to why D did not produce the expected results in this study may be the fact that, as suggested by Van Hout and Vermeer (2007), TTR can sometimes be a better measure than D in terms of concurrent validity (see Chapter 2). Another possible reason could be the fact that I used written data (by only female learners) for my analysis and, according to G Yu (2009), D seems to be a better predictor of speaking than writing data.

Comparing Study 1 and Study 2

Study 1 looked at both oral and written data, whereas the larger scale Study 2 only focused on written data. The ratings that we had tried to predict in Study 1 were only the IELTS overall ratings, following the assumption that this would not differ from the analytic ratings for each category (due to the halo effect). However, in the second study (Study 2) we have tried to predict not only the overall holistic rating but also the lexical rating (given for the writing IELTS band descriptor as one of the analytic traits). This made sense since the measures used for predicting the IELTS ratings are all measures of vocabulary knowledge, therefore we wanted to investigate whether the model for the lexical ratings could have better predictive validity than the model for the overall rating (holistic rating).

In Study 1, Guiraud was the variable that could explain 36.8% of the variability in the overall holistic oral scores. In Study 2, it was also found to be one of the predictors in the model in combination with two others, TTR and Guiraud advanced and they could explain 49.2% of the variability in the holistic ratings. Study 1 looked at written and oral data, Study 2 only at written data. My results show that Guiraud is a good predictor

for both, written and oral data. For the written data in Study 1, P_Lex and tokens were the two variables that could explain 22.4 % of the variability in the written overall (holistic) ratings. This was not the case for Study 2, which revealed that for the written holistic ratings the best predictors were TTR, Guiraud and Guiraud Advanced which accounted for 49.2% of the variability in the ratings. In Study 2 P_Lex was found to predict the lexical ratings along with TTR and Guiraud (whereas in Study 1 it was found to predict the holistic ratings). These three variables could explain 52.8% of the variability in the lexical ratings.

As can be seen, the results from Study 1 differed from those of Study 2. The model in Study 2 for the holistic scores is definitely improved over the one in the first study, because in Study one it explained 22.4% of the variance in the ratings whereas in Study 2 it explains almost 50 per cent (49.2%). I believe that the IELTS overall (holistic) score model's predictive validity is quite high at almost 50%. This means that almost half of the variance in the ratings can only be explained by measures of lexical richness, and this is an indication of the importance of vocabulary in teacher ratings. It seems that vocabulary plays a major role in second language testing and assessment. Furthermore, regarding the lexical rating model (in Study 2), the fact that the three measures of lexical richness (TTR, Guiraud and P_Lex) explain 52.8% of the variance in the ratings shows that, even though there are very good valid measures of lexical richness, vocabulary is indeed multidimensional, and other aspects can be added to the model to try to further explain the variance in the ratings.

Can we accept the hypothesis 'The model comprised of lexical richness measures and phrasal expressions count (formulaic count) will have predictive ability'.

Following these results, I have to reject my hypothesis as it was only partly confirmed. I hypothesised that the model's predictive ability would increase with the addition of the formulaic count variable, but it appears that this was not proved on this occasion. It seems that a more traditional measure such as the TTR proved to be a better predictor of ratings than the formulaic count. This result was in accordance with the results from a previous study by Fatahipour (2012). In addition, regarding examiners comments when awarding a specific band for Lexical Resource (from the rating scales) Brown (2003) shows that the use of formulae (learned expressions) can sometimes (not always)

work against some candidates. Some examiners seemed to view the use of formulaic sequences as a negative aspect regarding the candidates' vocabulary range or use and not something that shows greater proficiency (as is suggested by literature- see Chapter 4). Brown (2003:53) stated that 'although there were not a huge number of references to learned expressions or formulae, examiners typically viewed their use as evidence of vocabulary limitations'.

However, I still believe that the formulaic count could play an important role in improving the models' validity in the future by approaching it from a different angle. My study was mostly quantitative as, even though I included measures of lexical sophistication in the research, I was dealing mostly with numbers, ratios and frequencies. The next step should be to look at the formulaic count more qualitatively. For example, instead of just having a count of how many expressions were used (from Martinez and Schmitt's list), there should be a list of which ones were used and how many times. For instance, I do believe that if a student that had a formulaic count of 15 achieved a lower rating than someone else that had a count of 5, this could be because the first one just used 3 types (but repeated them) and the other could have used 5 different ones. After the unexpected results, I did go back to the data and chose two student cases that simulated the example above. It was indeed true that the student with the higher rating used a variety of formulaic expressions, whereas the other just repeated the same ones and achieved a lower grade. After this realisation I decided to go back to the data and conduct a small qualitative analysis which could reveal the reasons behind the surprising results.

6.4 QUALITATIVE ASPECT

Quantitative research did not exactly show the effect that formulaic sequences are believed to have on the final ratings on essays, so it was decided that a sample of essays would be selected and analysed further in an attempt to determine the factor that affects examiners' decision to give high or low ratings.

Therefore, we returned to the data after the original quantitative analysis. As Turlik states (2008:30) 'Quality can be an emotive term therefore researchers feel the need to *quantify the qualitative*'. This is exactly what I do in this part of the study and how the

term 'qualitative' is used in the thesis (use it to refer to qualitative aspects which are then quantified). The fact that I am counting the instances of formulaic language does not make them stop being an aspect of depth of knowledge (qualitative aspect). 30 essays were selected from the sample. All the essays with high ratings (that were rated with grade 8 and 7) were selected. There were only two essays rated 8 and these were included. In addition five essays which were awarded a rate of 7, twenty one essays that were awarded a 6, one essay with rate 4 and a rate 2 essay were randomly selected. This range was included to check the use of formulaic sequences and check why their use did not correlate with the teacher ratings. The ratings are holistic ratings either by Rater 1 or Rater 2 (the selection was random). An analysis was conducted to investigate the main characteristics of each category, the use of formulaic sequences, and the extent of their use. The following table shows the main characteristics of these essays.

Table 6.23: Summary of Essay Characteristics and Ratings

Essay number	Ratings	Tokens	Types	D	Formulaic Count
10.4	8	356	154	55.78	7
8.4	8	241	111	59.41	2
2.6	7	466	173	60.66	9
28.6	7	422	197	72.90	8
28.7	7	529	227	92.92	15
23.5	7	475	153	47.31	12
25.5	7	342	160	65.79	8
38.5	6	248	108	48.61	8
18.2	6	364	132	45.58	6
18.3	6	331	107	39.08	12
21.3	6	380	151	47.44	4
37.3	6	227	104	52.37	7
37.6	6	164	351	99.57	12
44.4	6	372	156	82.96	10
4.6	6	364	124	58.49	10
27.6	6	544	149	58.29	10
36.5	6	373	156	47.30	8
44.5	6	422	165	59.98	8
19.7	6	409	159	60.95	6
25.7	6	306	122	72.52	5
36.7	6	342	135	52.14	7
43.7	6	457	181	78.88	4
38.3	6	236	95	44.72	4
17.4	6	249	136	79.70	7
21.6	6	410	148	48.96	9
36.5	6	373	156	47.30	8

29.6	6	399	155	88.06	8
29.7	6	424	161	91.24	10
27.2	4	559	140	38.18	17
34.1	2	123	62	42.12	1

It seems that the numbers in the last column of the table above (formulaic count) cannot justify the rating given according to our hypothesis (people with higher use of collocations obtain higher marks). Therefore, it is not necessarily true that the more collocations a learner uses, the higher the mark they achieve. Why does this happen?

A detailed description of not only how many formulaic sequences were used by each learner, but also which ones and with how many repetitions was prepared in order to help us further understand collocational use. This detailed description of the formulaic sequences used (from Martinez and Schmitt's list) can be found in Appendix 13.Table 6.24 below summarises the use of each and every one of these instances.

On the left hand side of the table all the formulaic sequences used are listed, and on the top is the number of each essay. For each essay you can see the total of formulaic sequences used and the given rating. In addition, the last column (right hand side) reveals the total number of each sequence used in the sample.

On the next page is a table that shows all occurrences of formulaic sequences in this 30 essay sample.

Table 6.24: Summary of Formulaic Use and Ratings

Form Seq	2.6	4.6	8.4	10.4	17.4	18.2	18.3	19.7	21.3	21.6	23.5	25.5	25.7	27.6	27.2	27.7	28.6	28.7	29.6	29.7	34.1	36.5	36.7	37.3	37.6	38.3	38.5	43.7	44.4	44.5	Total
there are	3	2	1	2	2	1	5	1		2	3	1	1		5	1			2		1	3	1	2	1	2	1		2	2	47
based on				1																											1
instead of				2			1				1																	1			5
all over				1							1		1																		3
out of				1				1																							2
cause			1		2	2	4					1			1									4		2	2		1	3	23
carry out	1																														1
deal with	2										1								4	3		1		1							12
each other	1																	1		2											4
lead to	1																	2													3
come up with	1																														1
as well as		1																													1
is to								1								1															2
such as				2				1		1						3									1						8
even though																1			1												2
in full																1															1
by way of																1															1
a lot											1				6	1													1		9
a number of																1															1
as well																1															1
tend to																1															1
that is																1		1							2						4
get to																2															2
one another																1															1
a good		2											3			1									1				1		8
contrary to																1															1
over time																1															1
can tell																1															1
going to		1					2			2	1														1						7
have to				1							1	1													1		1			2	7
is to											1																				1
look for											1																				1
in my opinion		1									1																				2
think about												1											1								2
at the same time												1																			1
take care of										1		2													1					2	6
at work										1		1					8								3	1					14
there is																			1						1	1	2				5
too much		1																								1					2
for all					1					1					1											1					4
kind of					1				1						1					1					2						6
put it					1																										1
to go								2	1																						3
more and more								1																							1
I mean																							2								2
a little																													1		1
focus on																													1	1	2
take advantage																													1		1
in addition		1																				1									2
think it is		1																													1
find it																	1														1
in the end																						1									1
work on																														1	1
used to						2																									2
in fact																							1								1
so that																							1								1
provide for																							1								1
good at																			1												1
you see																				1											1
to me																				1											1
of course															3																3
or anything																				1											1
Total	9	10	2	7	7	6	12	6	4	9	12	8	5	10	17	1	8	15	8	10	1	8	7	7	12	4	8	4	10	8	
Grade	7	6	8	8	6	6	6	6	6	6	7	7	6	6	4	6	7	7	6	6	2	6	6	6	6	6	6	6	6	6	

The table above shows that some essays achieved a higher grade even though their formulaic count was low (or achieved low grades even though their formulaic count was high). There are many interesting examples (all the examples discussed below were colour coded). These examples are considered important for the topic because they could explain the absence of the formulaic count variable from our IELTS ratings predictive model. For instance, if you look at the total number of formulaic sequences and the grade given (at the bottom of the table), one can see that even though some essays shared the same number of sequences, they obtained different grades. In fact, some essays with a high number of sequences were graded lower than others with a smaller number of sequences. This was further investigated and it was revealed that some essays had many repetitions of the same sequences. For example, essays 23.5 and 18.3 both included 12 formulaic sequences. However, the first was awarded a grade 7 and the second one (18.3) a grade 6.

Essay 23.5

Total number of formulaic sequences: 12

a lot (1)

going to (1)

have to (1)

is to (1)

there are (3)

deal with (1)

look for (1)

instead of (1)

all over (1)

in my opinion (1)

Essay 18.3

Total number of formulaic sequences: 12

going to (2)

there are (5)

instead of (1)

cause (4)

In the first essay there were 9 types (with one type being repeated 3 times, therefore giving a total number of formulaic sequences of 12), whereas in the second essay there were only 4 types that were repeated throughout the essay. This could be a possible explanation as to why the essays did not receive similar marks.

This can also be observed when someone compares essays 27.2 and 28.7. These two essays had the highest number of formulaic sequences from the sample. Shown below are the formulaic sequences for each essay:

Essay 27.2

Total number of formulaic sequences: 17

a lot (6)

of course (3)

there are (5)

kind of (1)

for all (1)

cause (1)

Essay 28.7

Total number of formulaic sequences: 15

a lot (1)

a number of (1)

as well (1)

lead to (2)

tend to (1)

even though (1)

that is (1)

get to (2)

one another (1)

a good (1)

contrary to (1)

over time (1)

can tell (1)

The first essay includes 17 instances of formulaic language use but only received a grade 4. The second essay has 15 formulaic sequences but was awarded a grade 7. After further investigation it was revealed that the first essay had many repetitions (only 6 formulaic sequences, and three of them were repeated 6, 3 and 5 times). The second essay had 15 sequences, and only 2 of them were repeated (therefore there were 13 types).

All the previous examples could be possible explanations for obtaining different results from what was originally expected and was one of the hypotheses (that the higher the grade, the greater the use of formulaic sequences in a user's essay). This could be the reason why the formulaic count could not be used as one of the predictive variables in my ratings model. This may be due to the nature of the variable (I only counted formulaic sequences at that stage and did not check which ones were used or how many times they were repeated). One of the only essays that seemed to fit the expected results was essay 34.1, which can be considered as a good example of what was expected at the beginning. This essay received the lowest mark (grade 2), and was the only essay in the sample that had 1 formulaic sequence.

Essay 34.1

Total number of formulaic sequences: 1

there are (1)

However, there are some remarkable cases that need to be pointed out and discussed. For example, essay 8.4, which was awarded the highest grade (grade 8) only included 2 instances of formulaic language.

Essay 8.4

Total number of formulaic sequences: 2

there are (1)

cause (1)

In order to obtain more answers and improve the current model of predicting ratings based on lexical richness, further research is needed. The next step would be to qualitatively analyse all the essays for my research to try and answer all the questions that arose from this 30 essay sample work. Maybe other factors that could be affecting the grade, such as the topic of the essay should be investigated. This is definitely something worth looking into in the future.

The analysis so far has concerned the relationship between the number of formulaic sequences used by each participant and the ratings (given grades). However, we can look at the results of this qualitative research from another perspective, such as the effect of learning and teaching formulaic sequence. In order to do this we can isolate two columns from Table 6.25: the formulaic sequences, and their total number used in all 30 essays.

Table 6.25: Total Number of Formulaic Sequences Used in Sample (30 essays)

there are	47		more and more	1
based on	1		I mean	2
instead of	5		a little	1
all over	3		focus on	2
out of	2		take advantage	1
cause	23		in addition	2
carry out	1		think it is	1
deal with	12		find it	1
each other	4		in the end	1
lead to	3		work on	1
come up with	1		used to	2
as well as	1		in fact	1
is to	2		so that	1
such as	8		provide for	1
even though	2		good at	1
in full	1		you see	1
by way of	1		to me	1
a lot	9		of course	3
a number of	1		or anything	1
as well	1			
tend to	1			
that is	4			
get to	2			
one another	1			
a good	8			
contrary to	1			
over time	1			
can tell	1			
going to	7			
have to	7			
is to	1			
look for	1			
in my opinion	2			
think about	2			
at the same time	1			
take care of	6			
at work	14			
there is	5			
too much	2			
for all	4			
kind of	6			
put it	1			
to go	3			

We can see from the above list that out of the possible 505 phrasal expressions (formulaic sequences) from the list used in this study, only 63 formulaic sequences were found. There were some that were used far more than others. It seems that the most frequent and commonly used sequence is *there are,* which was used 47 times in total (in our 30 essay sample). The second most frequent was *cause* (23 times), and the third was *at work* (14 times). This list summarises the use of the sequences in this particular school in Dubai, and these results could be the effect of teaching in this particular school. Perhaps these sequences are taught more than others, and this could be something that the teachers could work on in order to improve the teaching of formulaic sequences. Maybe we can work more on the remaining sequences from the list to ensure that our students learn and use more than just a few basic and common ones. Again, this is something very interesting that needs to be investigated in another project in the near future.

6.5 ISSUES- LIMITATIONS OF THE STUDY

In this section, issues connected with research methodology that arise from the study are discussed. After the completion of Study 1 (pilot study), I suggested that more variables (apart from measures of vocabulary) could be added to improve the predictive validity of the IELTS model. Therefore, I suggested that in further research, variables such as 'number of years learning English' could be added for an improved version of my model. However, a limitation of Study 2 was that I could not use ethnographic data because the students were not asked to provide such information as it would be considered inappropriate due to cultural restrictions. Findings from previous research suggest that even if that variable was added, the model's predictive validity could remain stable. According to Green (2005), a variable that does not seem to correlate with writing scores/ratings is 'years of English'. Green (2005:54) states that: 'Years of English study was found not to correlate with writing scores and was excluded from the model'. There were also other problems/implications that should be noted regarding this study. First of all, the fact that the inter-rater reliability (especially for the holistic ratings) was low could be proven to be a hindrance. Furthermore, the topic variable was not controlled and this could have affected the results. According to Schoonen (2005:3), 'in writing assessment numerous sources of variance other than the writing

ability of the students contribute to the variance in writing scores' such as topic, discourse mode, time limits etc.

Regarding the statistical methods that were employed, some critical arguments have been made over the years by other researchers. Zareva (2005:547) supports the use of multiple regression analysis for predictive purposes:

> 'Multiple regression analysis, which is frequently employed for predictive purposes, and all possible regressions, which is one of several procedures used for identifying the most efficient predictors from a pool of variables'.

However, Barkaoui (2010) suggests that researchers should be careful with the use of such statistical methods.

Another limitation of the study could be the fact that, even though the essays used in the research were marked by trained IELTS examiners using the IELTS Writing Band descriptors, they were just academic essays produced in examination conditions, not IELTS essays as such. Moore and Morton's study (2005) showed that the IELTS Academic Writing Task 2 is similar to an academic essay, however, there are also important differences, which could cause differences in the prediction results. At the research design stage of this study it was presumed that access to IELTS data would have been granted, as was the case in previous studies. For example, Read and Nation (2002) used actual IELTS data (from actual IELTS tests) which included a wide range of band scores (from band score 4 to band score 8). It therefore seemed reasonable to hope that permission would be granted this time, which would have guaranteed a wider range of band scores, and benefitted the model. However, permission was denied.

A major and expected complication is that of the use of word lists. For the study many word lists were used for the lexical richness (lexical sophistication) calculations and formulaic count. There are researchers such as Möbarg (1997) who criticise vocabulary lists by arguing that any list, no matter how sophisticated it can be, is still not the real vocabulary of a language, but just a sample of it. Therefore, the real language used by the learner could not be captured by the use of these lists. Such issues are grounds for stimulating much future research. Furthermore, it has to be noted that the list used in Study 2 (PHRASE list) for researching formulaic language use was based on the BNC.

The BNC consists of 90% written texts of many different kinds and only 10% of transcribed speech. However, this 10% represents around 10 million words, recorded both in formal and informal contexts. This could be something that affected my study's outcome because my analysis was based on written academic essays. Therefore, these 10 million words taken from oral data and also from informal contexts could still be a lot. Thus, one could argue that the list could be more (or equally suitable) to be used in oral data analysis. However, the list could not be used in Study 1 (where oral transcriptions were analysed) because it did not exist when the analysis was conducted (2009). This could stimulate further research where the researcher could use the list with oral data and maybe texts of different kinds (not just academic texts).

Lastly, applying lexical richness measures to real-life teaching and testing situations may be problematic. In terms of testing, tests will never be the same as real and authentic (naturally occurring) language use. Even if researchers were able to acquire a large amount of language, it would be very time consuming to transcribe and analyse and would include manually counting formulaic language. This of course has implications about what each researcher would understand as formulaic language. Therefore it would be very hard for any automated electronic lexical measure to cope with this and it would only be possible with certain predetermined features which would never include the full richness of natural interaction.

6.6 COMPARISON BETWEEN TURLIK'S STUDY AND MINE

As was previously discussed in Chapter 5, the data used in Study 2 was data from an existing corpus collected by Turlik which was then made available to the public. In this section a comparison is provided between his findings and mine. I have to stress (as already mentioned in Chapter 5- Methodology section) that the data used (essays and IELTS ratings) were collected by Turlik but for my study a complete re-analysis of all the measures was conducted. For the data collection see Chapter 5- Methodology Section and Turlik's Chapter 5 (2008:128).

Turlik's study aim was to investigate, using 340 essays from EFL learners collected over a period of twenty-seven months (longitudinal study), how vocabulary growth could be modelled. He also wanted to check if trained EFL teachers and IELTS raters

would base their ratings on lexical richness. My study begins with the hypothesis that raters do base their ratings on lexical richness and further explores to what extent these ratings can be predicted by measures of lexical richness alone (what percentage of the given rating/judgement is based on vocabulary). One of the foundations of both studies is the definition used for lexical richness borrowed from Malvern et al. (2004) which assume that lexical diversity and lexical sophistication constitute lexical richness. Both studies (Turlik's and mine) investigated the lexical diversity and lexical sophistication of the texts using similar methods. It was not the purpose of his study to investigate or discuss arguments about which measure is the best and avoids to discuss or imply any competition between them. It was not the purpose of my study either to decide which lexical richness measure is the best but which (of the selected measures) are the most appropriate ones to include in a model that predicts IELTS ratings.

In terms of the measures used for the calculation of lexical richness in his study and mine, we both used number of types and number of tokens, D (as a measure of lexical diversity), P_Lex, Guiraud, and Guiraud Advanced. There were also other measures that were either used only by him or myself. He also used the use of advanced words (AWL), Limiting Relative Diversity and LFP whereas I used TTR and Formulaic count (operationalised using Martinez and Schmitt's List- see Methodology Section in chapter 5). He chose Guiraud, D and LRD as they claim to be the least affected by text length and P_Lex was included as an alternative of the LFP. He did not include Guiraud Advanced because as Malvern et al (2004) claim is sensitive to text length. However, Daller et al (2007) disagree and suggest that both Guiraud Advanced and Advanced TTR are valid measures and this explains the inclusion of Guiraud Advanced in my research.

Regarding lexical diversity, his results showed that the number of tokens and number of types as well as Guiraud and D increased significantly over the two years. In terms of lexical sophistication, the results revealed that Advanced Types (AWL), P_Lex, Guiraud Advanced, and Limiting Relative Diversity (LRD) increased significantly over the years. However, there was no clear evidence using the Lexical Frequency Profile (LFP) that during vocabulary growth basic vocabulary becomes less and is replaced by more advanced vocabulary (Turlik, 2008: 158).

The measures that were used in Turlik's research but not mine are the Advanced Types based on the AWL, Lexical Frequency Profile (LFP) and Limiting Relative Diversity (LRD).

His findings showed that all of the measures used (except from the LFP) increased significantly over the period of two years, therefore should definitely be found in my predictive model as well and account for the variability in teacher ratings. Regarding the lexical ratings by the IELTS raters there seemed to be an increase over the two years. There was also a significant increase in Holistic ratings over the two years. Something that needs to be taken into consideration regarding the results from both studies (mine and Turlik's) is the fact that 'the raters tended to award a mark reflecting writing proficiency at a certain level rather than an IELTS score, so the findings in respect of correlations of ratings (and regression analyses) should be viewed with this in mind' (Turlik 2008:162). Turlik also reported high correlations between the measures D, Guiraud, Guiraud Advanced and P_Lex (high correlations between D and Guiraud) and Guiraud Advanced and P_Lex. In addition, Inter-rater reliability was not high (page 169). After Turlik's regression analyses he concluded that types and/or tokens are very important in lexical and holistic ratings (page 179.) His main findings were that there is a general group and individual pattern of vocabulary development over time (essays written later at time will receive higher marks than earlier ones). There was an increase of sophisticated vocabulary (more complex vocabulary) as there was an increase of the AWL types but not a decrease of general vocabulary. There was a high correlation between D and Guiraud suggesting a very high relationship between the two (in this study). Guiraud Advanced, AWL and LRD all correlated therefore suggesting that the use of advanced words seemed to increase over time. One of the main differences in our studies is the fact that I used TTR as a measure of lexical diversity which proved to be one of the best variables to predict IELTS ratings. Turlik believed that TTR should not be used because there was an increase in text length between the first and seventh essay (the beginning and end of data collection).However, his study was longitudinal and its aim was to look at the lexical development (vocabulary growth) over the years. I treated the data differently (in terms of analysis) and one of the main reasons for the inclusion of TTR was because, despite its flaws (as already discussed in previous chapters), it is a very widely used measure and I wanted to see how it would perform in the creation of the predictive model.

CHAPTER 7- TESTING THE MODEL

7.1 INTRODUCTION

In this chapter a new study (Study 3) is introduced which tests my model against some new other data. The data used are the sample essays (actual IELTS data) provided by the IELTS Organisation on their website and can be found in the following address: http://www.ielts.org/PDF/113313_AC_sample_scripts.pdf. There are eight examples of essays and rater marks and comments. These are all Academic Writing Tasks. Even though the sample used is too small for me to be able to make any strong statistical analyses, I evaluate the performance of my model and include some qualitative analysis of the raters' comments (for example, some mention 'words').

7.2 TESTING THE MODEL USING STATISTICAL ANALYSIS (QUANTITATIVE APPROACH)

According to my predictive model from Study 2, holistic ratings can be predicted by two lexical diversity measures (TTR and Guiraud), and a lexical sophistication measure, Guiraud Advanced. A model consisting of these three measures can explain 49.2% of the variability in the holistic ratings (see the model explained in Chapter 6).

IELTS Holistic Rating= 3.553-4.495*TTR +0.304*Guiraud+ 0.540*Guiraud Advanced.

In order for the developed model to be tested against new, unseen, data, essays provided by IELTS through their website were used. For each of the eight essays, the measures of lexical diversity and sophistication were calculated including the necessary model inputs TTR, Guiraud and Guiraud Advanced metrics and can be found in the following table along with the IELTS Score:

Table 7.1: Calculations of Measures of Lexical Diversity for each script

Script number	Types	Tokens	TTR	Guiraud	P_Lex (lambda)	Guiraud Advanced	D	IELTS Score
1AA	56	131	0.427	4.893	1.3	0.173	30.09	5
1AB	58	165	0.352	4.515	2.36	0.571	26.71	6
1BA	74	177	0.418	5.562	1.88	0.651	36.47	6
1BB	81	192	0.422	5.846	2.13	1.032	41.2	7
2AA	109	206	0.529	7.594	1.13	1.120	67.02	5
2AB	139	271	0.513	8.444	1.21	1.637	67.52	6
2BA	122	228	0.535	8.080	2.12	1.384	76.7	5
2BB	150	349	0.430	8.029	2.12	1.232	62.08	7

It should be noted here that the scores of these eight essays range between 5 and 7. This is an important aspect of this dataset given the fact that in the dataset the holistic ratings' model was estimated on, only 13% of the observations had an average rating of 5 or more.

The relevance and importance of this finding became apparent as soon as the model was used to predict the IELTS Score. The point predictions with associated 95% Prediction Intervals ('prediction' rather than 'confidence' interval, given that the interval refers to predictions made on unseen data to the model) are included in the following table and have been calculated using the expression:

IELTS Holistic Rating Prediction = 3.553 − 4.495*TTR +0.304*Guiraud+ 0.540*Guiraud Advanced

Table 7.2: Point predictions for new scripts with associated 95% prediction limits

Script number	IELTS Score	Prediction	Lower	Upper
1AA	5	3.21	2.14	4.28
1AB	6	3.65	2.57	4.73
1BA	6	3.71	2.65	4.78
1BB	7	3.99	2.91	5.07
2AA	5	4.09	3.02	5.16
2AB	6	4.70	3.61	5.78
2BA	5	4.35	3.27	5.43
2BB	7	4.73	3.65	5.80

Based on the above table, as an example, the model estimate for the IELTS Score of Script number 1AA is 3.21 with a 95% prediction interval between 2.14 and 4.28. The actual score however was 5 which falls outside the range of values of the prediction interval. This is the case in 6 out of 8 essays, which is rather surprising given that the prediction intervals are constructed in such a way as to account for the variability of the data and therefore of any predictions. Only the actual IELTS scores for scripts 2AA and 2BA fell inside the corresponding prediction intervals.

The consistent underestimating of the IELTS Score by the model has made further investigation of the data in question, necessary. Any statistical model is built based on the data it was provided with and therefore will not perform adequately outside the range of values it was developed on. As a result and given the fact that only a small fraction of the initial essays were rated with scores that corresponded to the scores of the testing dataset (i.e. scores of 5 to 7), the exhibited underestimation of the IELTS Scores in the testing dataset was not that surprising.

To check this in more depth, for each of the 8 testing essays, a subset of the original essays was selected based on the proximity of the TTR, Guiraud and Guiraud Advanced metrics and the average score based on the two raters was calculated. The reason for doing this was to compare how similar essays were rated in the original dataset and whether the testing dataset ratings were higher than those. Taking essay 1AA as an example, its TTR, Guiraud and Guiraud Advanced metrics (values) were 0.427, 4.893 and 0.173 respectively as shown in the table below.

Table 7.3: Calculations of measures of Lexical Diversity for Script 1AA

Script number	Types	Tokens	TTR	Guiraud	P_Lex (lambda)	Guiraud Advanced	D	IELTS Score
1AA	56	131	0.427	4.893	1.3	0.173	30.09	5

The original dataset was interrogated to find essays with similar TTR, Guiraud and Guiraud Advanced values. More specifically, the selected essays from the original dataset, had values for the three metrics that fell within 1 standard deviation (SD) of the highlighted values in Table 7.3 (i.e. the test essay's corresponding metrics values). A table with the standard deviation of the 3 relevant metrics is given below:

Table 7.4: Standard Deviation of TTR Guiraud and Guiraud Advance in original data

	TTR	Guiraud	Guiraud Advanced
Standard Deviation	0.07	1.03	0.28

Hence, in the case of essay 1AA, essays with values satisfying the following expressions were selected:

0.427 -0.07 (i.e. 1 SD below) = 0.357 < TTR < 0.497= 0.427+0.07 (i.e. 1 SD above)

3.863 < Guiraud < 5.923

0 < Guiraud Advanced < 0.453

The average score for these essays was calculated and recorded. The results are shown in the following table:

Table 7.5: Comparison between predicted ratings and the original dataset

Script number	IELTS Score	Prediction	Lower	Upper	Similar Essays Average
1AA	5	3.21	2.14	4.28	3.25
1AB	6	3.65	2.57	4.73	3.63
1BA	6	3.71	2.65	4.78	3.73
1BB	7	3.99	2.91	5.07	4.13
2AA	5	4.09	3.02	5.16	3.81
2AB	6	4.70	3.61	5.78	5.00
2BA	5	4.35	3.27	5.43	4.00
2BB	7	4.73	3.65	5.80	4.38

The results suggest that even though there is significant underestimation of the model predictions compared to the IELTS Score, the model predictions are in fact comparable to the scores exhibited by the actual data in the original dataset (Column 'Similar Essays Average'). This finding would reinforce the notion that the 8 test essays are not representative of the essays used to build the statistical model in the first place and as such should not be used to assess its validity. Therefore, it would be a major advantage for a similar study to use and analyse real IELTS data as the basis for the model

(something which was originally intended here but was not possible in the end and proved to be a hindrance to the validity of the study.)

In addition, we could maybe explain some of the differences in the data (explain the scores) if some qualitative analysis of the examiners' comments is conducted.

7.3 QUALITATIVE ANALYSIS OF THE DATA

In this thesis it was considered important to include a qualitative analysis of my data because there may be differences (qualitative) that were not captured by the quantitative analysis and may be revealed only by using a more qualitative approach.

Following the example of Mayor et al. (2002) and Read and Nation (2002) I decided to conduct an exploratory qualitative analysis using the examiners' comments of this small number of scripts available.

7.3.1 The essays

There are four Academic Writing Task 1 essays and four Academic Writing Task 2 essays (for a full description of these tasks please refer to the IELTS section in Chapter 2- Section 2.8). In order to conduct a more qualitative analysis I will first explain how the essays are coded: All four Academic Writing Task A essays have the number 1 (1AA, 1AB, 1BA and 1BB) and there are two different answers A and B regarding each essay. Therefore, 1AA and 1AB is the same task answered by two different candidates then essays 1BA and 1BB refer to a different question and are the answers of two other candidates. The same applies to the other four Academic Writing Task B essays (coded 2AA, 2AB, 2BA and 2BB). Thus, essays 2AA and 2AB are essays answering the same question attempted by two different candidates and having been awarded different scores (band levels). The same applies to essays 2BA and 2BB. Subsequently, it would make sense not to only analyse the data set as a whole or each essay individually but to also provide comparative analysis for each set of two essays. All essays were transcribed into CHAT (to be analysed using CLAN) and text format (for Guiraud Advanced calculations). Please see Appendices for the transcribed essays (Appendix 14) and examiner comments (Appendix 15).

7.3.2 Examiners' comments

In order to understand the way raters think to arrive to a specific decision about a band mark, we should take a closer look at their comments. I will first provide a summary of the examiners' comments and then compare them using each set of essays that are under the same topic (as previously explained). The analysis of the qualitative data took a top down perspective and looked for instances where 'vocabulary' and relevant terms such as 'length of answer' or adjectives such as 'limited' or 'varied'-when referring to vocabulary use- were used to justify a negative or positive mark. All these instances were highlighted in the excerpts below.

First, below is a comparison between examiners' comments for essays 1AA and 1AB:

ESSAY 1AA

Examiner comment
Band 5
*'The **length of the answer** is just acceptable. There is a good attempt to describe the overall trends but the content would have been greatly improved if the candidate had included some reference to the figures given on the graph. Without these, the reader is lacking some important information. The answer is quite difficult to follow and there are some punctuation errors that cause confusion. The structures are fairly simple and efforts to produce more complex sentences are not successful.'*

The examiner does not mention vocabulary as such but comments on the length of the essay (number of tokens) as '*just acceptable*'. Examiner also comments on errors (punctuation errors) and grammar (simple, not complex sentences). The lack of complex sentences is also relevant to vocabulary because vocabulary is embedded in other language aspects such as grammar.

ESSAY 1AB

Examiner comment
Band 6
'The candidate has made a good attempt to describe the graphs looking at global trends and more detailed figures. There is, however, some information missing and the information is inaccurate in minor areas. The answer flows quite smoothly although connectives are overused or inappropriate, and some of the points do not link up well. The grammatical accuracy is quite good and the language used to describe the trends is wellhandled. However, there are problems with expression and the appropriate choice of words and whilst there is good structural control, the complexity and variation in the sentences are limited.'

The examiner comments on content, cohesion and coherence and grammatical accuracy. They also mention problems with appropriate choice of words and complexity and variation in the sentences.

It seems that the essay receiving the lower mark here (band 5) was the one in which the examiner commented negatively on vocabulary, and more specifically on the number of tokens (the length of the essay). If we look at the opening sentences from both essays (see below) one can see the difference in vocabulary use. There are more sophisticated, less frequent words in the second essay which was placed at band level 6. Here are the opening sentences of Essay 1AA and 1AB:

Opening sentence from Essay 1AA

'This is a bar chart of the number of men and women in further education in Britain in three periods.'

Opening sentence from Essay 1AB

'According to this graph, the number of men and women in further education in Britain shows the following patterns.'

The influence of vocabulary on rater judgement can also be stressed from the following examiners' comments example. Here is the comparison between the examiners comments on essays 1BA ad 1BB:

ESSAY 1BA

Examiner comment
Band 6
'The answer has an appropriate introduction which the candidate has attempted to express in his/her own words. There is good coverage of the data and a brief reference to contrasting trends. The answer can be followed although it is rather repetitive and cohesive devices are overused. In order to gain a higher mark for content, the candidate would be expected to select the salient features of the graph and comment primarily on these. Sentences are long but lack complexity. There are some errors in tense, verb form and spelling which interfere slightly with the flow of the answer.'

The examiner mainly comments on grammar. There is no specific reference to vocabulary (apart from spelling) justifying the given mark/rating.

ESSAY 1BB

Examiner comment
Band 7
*'The answer deals well with both the individual media trends and the overall comparison of these trends. The opening could be more fully developed with the inclusion of information relating to the groups studied and the period of time during which the study took place. There is a good variety of cohesive devices and the message can be followed quite easily although the expression is sometimes a little clumsy. Structures are complex and **vocabulary is varied** but there are errors in word forms, tense and voice though these do not impede communication.'*

The examiner comments on vocabulary and states that 'vocabulary is varied' and even though there are some errors they do not impede communication.

The essay that receives the highest mark between the two in this set is the essay that received comments on vocabulary, and especially lexical variation. It is quite remarkable that one of the aspects that were taken into account when rating on of the higher band essays was lexical variation/diversity.

This variation is highlighted in the following extracts from essays 1BA and 1BB below.

Extract from Essay 1BA

*'The graph shows the percentage of audiences over 4 years old of UK follows the radio and television throughout the day during the period October December 1992. It has been observed from the graph that less than 10% audiences follows the radio at 6:00 am and the percentage **raised** to a pick around 30% at 8 am and decline gradually to around 10% during the period 2:00 to 4:00 pm and again **raised** a bit to around 12% between 4:00 to 6:00 pm then again **dropped** to below 10% at around 10 pm. The rate again **raised** to a bit between 10:00 pm to 12:00 pm and then **dropped** slowly by 4:00 am.'*

Extract from Essay 1BB

*'The bold graph shows the television audiences throughout the day. It shows that the percentage of audiences is zero percent in early morning but it **gradually rises up** to ten percent at 8:00 am and maintains the same for the next two hours. There is a **slight fall** in percentage in next two hours however after that it **rises sharp** up to twenty percent within the next two hours. After this the graph rises very fast and **attains its' peak** at 10 pm which is about forty five percent. The graph **gradually falls down** and at 2:00 am it is at five percent. The thinner graph shows the percentage of radio audiences. Unlike the television one the*

peak percentage of the radio audiences is at 8:00 am which is about 30 percent. Then it gradually falls and it corresponds with the television one at two pm.'

One can see that there is a repetition of certain vocabulary items in the first extract (*rise, dropped)* whereas, there is more diversity (use of different words) in the second extract which received the higher mark/rating. This essay (and examiner comments for this specific essay) provides more evidence to the argument that vocabulary plays a major role in teacher ratings.

Below is the comparison between essays 2AA and 2AB. Essay 2AA received a quite low rating and provides an example of how vocabulary use can negatively influence the examiners' ratings.

ESSAY 2AA

Examiner comment
Band 5
*'**The answer is short** at just over 200 words and **thus loses marks** for content. There are some relevant arguments but these are not very well developed and become unclear in places. The organisation of the answer is evident through the use of fairly simple connectives but there are problems for the reader in that there are many missing words and word order is often incorrect. The structures are quite ambitious but often faulty and **vocabulary is kept quite simple.***'

One of the factors taken into account here was the number of tokens ('*answer is short...thus loses marks*'). In addition, the examiner seems to be influenced by the use of vocabulary as they state that '*vocabulary is kept quite simple*' (lacking lexical sophistication- more sophisticated, less frequent words).

ESSAY 2AB

Examiner comment
Band 6
'There are quite a lot of ideas and while some of these are supported better than others, there is an overall coherence to the answer. The introduction is perhaps slightly long and more time could have been devoted to answering the question. The answer is fairly easy to follow and there is good punctuation. Organisational devices are evident although some areas of the answer become unclear and would benefit from more accurate use of connectives. There are some errors in the structures but there is also evidence of the production of complex sentence forms. Grammatical errors interfere slightly with comprehension.'

The examiner does not mention vocabulary in their comments.

It seems that the essay that receives the lowest mark (Band 5) from the two is the essay which received negative comments (regarding vocabulary) from the examiner. Particularly, the examiner mentions the number of tokens (length of essay) and the non-existence of more sophisticated, less frequent words – they state that vocabulary is quite simple.

Finally, below is the comparison between the final set of essays, essay 2BA and 2BB.

ESSAY 2BA

Examiner comment
Band 5
*'Although the script contains some good arguments, these are presented using poor structures and the answer is not very coherent. The candidate has a clear point of view but not all the supporting arguments are linked together well and sometimes ideas are left unfinished. There is quite **a lot of relevant vocabulary but this is not used skilfully** and sentences often have words missing or lapse into different styles. The answer is spoilt by grammatical errors and poor expression.'*

Essay 2BA received a quite low rating and, once again, vocabulary seems to play a negative role in the examiner's decision. Regarding essay 2BA, the examiner mentions incorrect use of vocabulary.

ESSAY 2BB

Examiner comment
Band 7
*'The answer is wellwritten and contains some good arguments. It does tend to repeat these arguments but the writer's point of view remains clear throughout. The message is easy to follow and ideas are arranged well with good use of cohesive devices. There are minor problems with coherence and at times the expression is clumsy and imprecise. There is a wide range of structures that are well handled with **only small problems in the use of vocabulary**, mainly in the areas of spelling and word choice.'*

Essay 2BB received a high rating. The examiner's comments which state that there are '*only small problems in the use of vocabulary*' show that examiners are indeed negatively or positively influenced (when rating IELTS written essays) by the (correct) use of vocabulary.

Again, in this data-set both examiners seem to comment on vocabulary use and the candidate's essay with less problems in the use of vocabulary (essay 2BB) receives a higher rating (Band 7).

An example of the lack of skillfulness regarding vocabulary use in essay 2BA is highlighted in the following extract:

Extract from Essay 2BA

'*Each country do not give threat to the country*. *Because they know if the country destroys cities, then other will create problems from them. So it is well-balanced and world peace maintains peacefully. Though **there are sometimes creates problems** by the nuclear technology but sometimes it also help the mankind in the field of medical and engineering sectors.*'

On the contrary, a small extract from essay 2BB (below) indicates that the candidate does not seem to face problems with vocabulary, resulting in the examiner's decision to place them on a higher band level (Band 7).

Extract from Essay 2BB

'*Nuclear power is an alternative source of energy which is carefully being evaluated during these times of energy problems. During these years we can say that we have energy problems but in more or less 50 years, we will be facing an energy crisis. Nuclear power is an alternative source of energy and unlike other sources such as solar energy, nuclear power is highly effective for industrial purposes. If it is handled correctly there really is no danger for the public.*'

7.3.3 Discussion of results from both quantitative (testing the model) and qualitative analysis (examiners' comments)

Regarding the quantitative analysis: it seemed that the model underestimated the IELTS scores and this was explained from the dataset chosen to base the model on. Therefore, I acknowledge the fact that the range of the ratings from the existing corpus (Turlik's corpus) used for Study 2 belong to a specific range (only 13% of the dataset were awarded Band 5 or above - the range is not wide) and this could be a potential flaw and a risk to the validity of the study and its findings. The model would of course perform better if it had been based on real IELTS data (essays and writings) and this is something to be pursued in the future.

It has to be noted here that in the two essays that the model seems to predict the band rating (Task 2AA and 2BA) vocabulary seemed to be one (if not the first) of the factors that influenced the raters decisions to award those specific ratings. In addition, both of these essays are Task 2 essays therefore similar to the essays that my model was based on.

One main difference between my data set and the new IELTS data is the fact that the 8 sample essays consisted of both Academic Writing Task A and Task B. As already discussed in Chapter 2 task choice may affect any research results as the vocabulary needed to successfully complete Task 1 is quite different in nature than Task 2. The data used for my study (Study 2), even though they were not actuals IELTS essays, were based on Task 2 questions. However, both tasks are assessed for their Lexical Resource.

The main finding from the qualitative analysis is its clear confirmation of the importance of vocabulary in language proficiency. This is revealed in the IELTS examiners' comments, the IELTS scores themselves (the better the vocabulary, the higher the score) and by the predictive model established in this research, where almost 50% of the variation in scores can be explained by measures of lexical richness alone. However, there are important aspects to investigate in order to improve the model, to explain the remaining 50% of variation in the ratings. Examiners comment on other aspects, such as grammatical accuracy and errors, but there is one feature that they all mention when assessing for Band 7, namely the use of cohesive devices and cohesion in general. This is something that could be investigated in the future as an idea for further research. Taking a mixed methods approach provides the researcher with the opportunity to complement the analysis. As shown here, bringing the two together allows for better insight in the data.

CHAPTER 8- CONCLUDING REMARKS AND SUGGESTIONS FOR FURTHER RESEARCH

In conclusion, this thesis was focused on vocabulary knowledge and its relationship with teacher ratings, particularly in IELTS tests. The first part of the thesis is concerned with discussing definitions of lexical knowledge and lexical richness, moving on to a discussion of the different dimensions of these two constructs, which are the principle focus of this study. Although historically there was a deficiency of methods for measuring vocabulary knowledge, now there is a plethora of such measures requiring a discussion of the merits and demerits of each in order to justify those finally chosen for use in this thesis.

The measures under investigation in this study were: number of tokens, TTR, D, Guiraud (measures of lexical diversity), and number of types, Guiraud Advanced, P_Lex (measures of lexical sophistication). All of these are measures of breadth (size) of vocabulary knowledge, and also measures of productive vocabulary knowledge. In the second study (Study 2) a measure of depth of vocabulary was added to the investigation. Vocabulary knowledge is undoubtedly an important aspect of language proficiency. There is a relationship between vocabulary richness and language proficiency, and various studies indicate that vocabulary knowledge could be used as a predictor of language proficiency. In particular, research shows that measures of lexical sophistication should have higher correlations with teacher ratings than measures of lexical diversity. With all these in mind, this thesis intended to answer the following research questions: Would the measures of lexical sophistication have higher correlation (than measures of lexical diversity) with the teacher ratings? To what extent could IELTS ratings be predicted with measures of vocabulary richness? In addition, Study 2 also investigated the extent to which a measure of depth of vocabulary knowledge (formulaic language-formulaic count) would improve the model's predictive validity.

Chapter 3 presented the first (pilot) study. The results for the written data showed that the variables with higher correlations with the written overall score were the types, Guiraud Advanced, and P_Lex. These are all measures of lexical sophistication, and this finding confirms the hypothesis that measures of lexical sophistication should

correlate higher with the ratings than measures of lexical diversity. The predictive model for the written data consisted of two variables: tokens and P_Lex, and could explain 22.4% of the variance in the written overall score. The model for predicting the oral overall score consisted of a single variable, Guiraud, which is a measure of lexical diversity even though lexical sophistication measures were expected to be in the model. A possible explanation for the results could be the nature of the different tasks. Written language is more formal, therefore more sophisticated words are used; oral language is more colloquial.

Chapter 6 presented the results of Study 2 and a comparative analysis between this study and Turlik (2008). There were correlations between all the measures of lexical richness (diversity and sophistication) and the teacher ratings. The lexical ratings model consists of three variables: two of lexical diversity, and one of lexical sophistication. TTR, Guiraud and P_Lex are the three measures (variables) that can explain 52.8% of the variability in the lexical ratings. In addition, holistic ratings can be predicted by the same two lexical diversity measures (TTR and Guiraud), but using Guiraud Advanced, a different measure of lexical sophistication. The model consisting of these three measures can explain 49.2% of the variability in the holistic ratings. It seems that for the formulaic count, even though there was a correlation with the teacher ratings, the correlation was not high. Therefore, it did not improve my model's predictive validity and had to be excluded from the model (being a non-statistically significant coefficient).

The results were quite surprising and very interesting. Firstly, and most surprisingly, D did not appear in any of the predictive models, despite having been empirically tested in many different linguistic fields and different languages (Treffers-Daller, 2013), which had suggested that it would be a good measure of lexical diversity for the IELTS model. In addition, many studies see it as a good predictor of proficiency (for example Crossley et al. (2011)), so the result was unexpected. Instead, TTR seemed to be a better predictor. The second unexpected finding was the fact that the formulaic count should have showed higher correlations with the teacher ratings and should have improved the model's predictive validity but it did not. A possible explanation could be that raters do not pay particular attention to formulaic sequences because they are not aware that this is an issue for the learners: they take them for granted. Speakers tend

not to be aware of their use of formulaic sequences in their L1, but rather process them unconsciously.

In Study 2, the instances of formulaic language (operationalised from Martinez and Schmitt's list of 505 most frequent phrasal expressions) were counted and added to the calculations. However, the data was not treated qualitatively at first. After the original results of Study 2, it seemed likely that the research findings would be enhanced if it was known not only how many phrases were used by each candidate but also which ones were used. After analysing the data from a more qualitative perspective, it was revealed that some essays with a high number of formulaic sequences that received low ratings (low band level scores) had many repetitions. In addition, the 'qualitative' analysis showed that only 63 out of the 505 possible phrasal expressions (formulaic sequences) were used in those 30 essays, which could indicate that some expressions are easier (compared to other expressions) for teachers to teach and students to learn first. The fact that this 'qualitative analysis' was conducted using a small sample of 30 essays can be considered as a promising start, and should definitely be addressed in further research. It seems highly likely that the model could be further improved and used in future research. As Xi suggests, 'Computer technologies will undoubtedly advance and become even more pervasive in our language learning and assessment practices' (2010:298), so this model could perhaps be used for automated scoring in the future. Study 3 (which is presented and discussed in Chapter 7) tested the existing predictive model using real IELTS data (essays) from the IELTS website.

The limitations of the present study need to be addressed in future studies. What proved to be the main hindrance in this research was the fact that the model was not based on real IELTS data (only tested on those in Study 3), as IELTS did not grant access to their database. If a researcher could replicate this study in the future using essays from the organisation's database, a fully functional model would be produced, which could be used for pedagogic purposes. This study however, contributes to the field by supplying the basis for the creation of a model which can be used as a tool, and by supporting or disproving certain arguments from past research. For example, the fact that TTR is considered a flawed measure was something we knew in the field. However, it seems that TTR is one of the best predictors for the IELTS ratings. It seems that its text length dependence flaw makes it a good predictor because the better texts are usually longer.

My research also raises several interesting questions. Why was an established measure such as D not in the model, even though recent literature argues that it is one of the best predictors of teacher ratings? Do we need to go back to basics (tokens and types and all the ratios based on them) instead of trying to invent new, more sophisticated measures? In the last 20 years there have been various attempts by researchers to either improve existing measures of lexical richness (diversity and sophistication) or develop and introduce new ones that would be better than others. The findings of this study state that almost 50% of the variance of IELTS Writing teacher ratings can be explained by using 3 measures, two of which are old, traditional, and have been heavily criticized, especially TTR. Does this mean that we should go back to basics? After all the research has been conducted, are the number of tokens and the number of types enough to measure lexical diversity?

Lastly and more importantly, the fact that the model in this study explains nearly 50% of the variance of IELTS Writing Scores confirms the argument that vocabulary is indeed one of the most important factors in language testing and assessment, and has a strong relationship with all language skills (Schmitt, 2010). Schmitt reports that findings from previous studies show that vocabulary accounts for 37-62% of the variance in proficiency scores. One example mentioned in the thesis is the study by Crossley et al. (2011b) whose findings revealed that lexical diversity could explain over 45% of the variation in human ratings in general, and in this particular case TOEFL scores. Even though the remaining variation between the scores could be further explained in the future by additional variables, it is quite remarkable that almost half of it is explained by a single aspect, vocabulary knowledge. This study supports the findings reported by Schmitt and confirms the following statement:

> 'Considering the multitude of the factors which could affect these scores (e.g. learner motivation, background knowledge, familiarity with test task), it is striking that a single factor, vocabulary knowledge, can account for such a large percentage of the variation.' (Schmitt, 2010:4)

REFERENCES

Adams, M. L. (1980). Five Co-occurring Factors in Speaking Proficiency. In J. Firth (ed.): *Measuring Spoken Proficiency*. Washington DC: Georgetown University Press, pp. 1–6.

Adolphs, S. and Schmitt, N. (2003) Lexical Coverage of Spoken Discourse. *Applied Linguistics*. 24 (4), pp. 425-438.

Aghbar, A. A. (1990 October) *Fixed Expressions in Written Texts: Implications for Assessing Writing Sophistication*. Paper presented at a meeting of the English Association of Pennsylvania State System Universities (ERIC Document Reproduction Service No. 352 808).

Akbarian, I. (2010) The Relationship Between Vocabulary Size and Depth for ESP/EAP Learners. *System*. 38 (3), pp. 391-401.

Albrechtsen, D., Haastrup, K., and Henriksen, B. (2008) *Vocabulary and Writing in a First and Second Language: Process and Development.* Basingstoke: Palgrave Macmillan.

Alderson, J.C. (1981) Report of the Discussion on Communicative Language Testing. In Alderson, J.C and Hughes, A. (Eds). *Issues in Language Testing. ELT Documents 111*. London: British Council

Alderson, J.C. (2005). *Diagnosing Foreign Language Proficiency*. London: Continuum.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In: A.P. Cowie (ed.): *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press, pp. 101-122.

Al-Zahrani, M. S. (1998) *Knowledge of English Lexical Collocations Among Male Saudi College Students Majoring in English at a Saudi university*. PhD, Indiana University of Pennsylvania.

Arnaud, P. J., & Savignon, S. J. (1997). Rare words, complex lexical units and the advanced learner. In J. Coady & T. Huckin (Eds.). *Second language vocabulary acquisition* Cambridge, UK: Cambridge University Press, pp. 157-173.

Bachman, L. (2000) Modern Language Testing in the Turn of the Century: Assuring that What we Count Counts. *Language Testing.* 17 (1), pp. 1-42.

Bahns, J., & Eldaw, M. (1993) Should we Teach EFL Students Collocations? *System* 21, pp. 101–114.

Banerjee, J., Franceschina F. and Smith, A.M. (2004) Documenting Features of Written Language Production Typical at Different IELTS Band Score Levels. *IELTS Research Reports 7,* pp. 241-309.

Barkaoui, K. (2010) Explaining ESL Essay Holistic Scores: A Multilevel Modeling Approach. *Language Testing*. 27 (4), pp. 515-535.

Beglar, D. (2010) A Rasch-based Validation of the Vocabulary Size Test. *Language Testing*. 27 (1), pp. 101-118.

Beglar, D. and Hunt, A. (1999) Revising and Validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*. 16 (2), pp. 131-162.

Benson, M. (1989) The Structure of the Collocational Dictionary. *International Journal of Lexicography*, 2, pp. 1–14.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, pp. 181-190.

Biber, D., S. Johansson, G. Leech, S. Conrad, and Finegan, E. (1999) *Longman Grammar of Spoken and Written English.* London: Longman.

Biber, D., Conrad, S., Reppen, R., Byrd, P. and Helt, M. (2002) Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly*. 36 (1), pp. 9-48.

Biber, D., Conrad, S. and Cortes, V. (2004) If you look at...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*. 25 (3), pp. 371-405.

Biskup, D. (1992) L1 Influence on Learners Renderings of English collocations. A Polish/German Empirical Study. In: P.J.L. Arnaud and H. Bejoint, (Eds): *Vocabulary and Applied Linguistics*. Basingstoke: Macmillan, pp. 85–93.

Bloomfield, L. (1933) *Language* .London: Allen and Unwin.

Blue M.G., Milton, J. and Saville, J. (2000) *Assessing English for Academic Purposes.* Oxford: Peter Lang, pp. 237-255.

Boers, F. (2000) Metaphor Awareness and Vocabulary Retention. *Applied Linguistics*. 21 (4), pp. 553-571.

Boers, F., Eyckmans, 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), pp. 245-261.

Boers, F. and Lindstromberg, S. (2009) *Optimizing a Lexical Approach to Instructed Second Language Acquisition.* London: Palgrave MacMillan.

Bogaards, P. (2000) Testing L2 Vocabulary Knowledge at a High Level: the Case of the Euralex French Tests. *Applied Linguistics*. 21 (4), pp. 490-516.

Bogaards, P. and Laufer, B. (2004) *Vocabulary in a Second Language: Selection, Acquisition, and Testing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bolander, M. (1989). Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. *Bilingualism across the lifespan: Aspects of acquisition, maturity, and loss*, pp. 73-86.

Bonk, W. J. (2000) *Testing ESL Learners' Knowledge of Collocations* (Report No. FL801 384). (ERIC Document Reproduction Service No. ED442309).

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, amt018.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*, *15*(1), pp. 45-85.

Broeder, P. & Voionmaa, K. (1985) General Procedures for Lemmatisation. In P.Broeder, G. Extra & R. Van Hout (1987) Measuring Lexical Richness and Variety in Second Language Use. *Polyglot* ISSN 0142-7202.

Brown, A. (2003) An Examination of the Rating Process in the Revised IELTS Speaking Test. *IELTS Research Reports 6,* pp. 41-70.

Brown, D. (2011) What Aspects of Vocabulary Knowledge do Textbooks Give Attention to? *Language Teaching Research*. 15 (1), pp. 83–97.

Bucks, R.S., Singh, S., Cuerden, J.M. and **Wilcock, G.K.** (2000) Analysis of Spontaneous, Conversational Speech in Dementia of Alzheimer Type: Evaluation of an Objective Technique for Analyzing Lexical Performance. *Aphasiology* 14, pp. 71–91.

Cameron, L. (2002) Measuring Vocabulary Size in English as an Additional Language. *Language Teaching Research*. 6 (2), pp.145-173.

Carroll, J.B. (1964) *Language and Thought*. Englewood Cliffs, NJ: Prentice Hall.

Carter, R. (1988) Vocabulary, Cloze and Discourse: An Applied Linguistic View. In R. Carter & M.McCarthy (Eds) *Vocabulary and Language Teaching.* Harlow: Longman, pp. 161-180.

Carter, R. (2000) *Vocabulary: Applied Linguistic Perspectives*. London: Routledge.

Chapelle, C. (1998) Construct Definition and Validity Inquiry in SLA research. In Bachman, L.F. and Cohen, A.D., editors, *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, pp. 32–70.

Chui, A.S.Y. (2006) A Study of the English Vocabulary Knowledge of University Students in Hong Kong, *Asian Journal of English Language Teaching,* 16, pp. 1–23.

Clark, H.H. (1970): Word Associations and Linguistic Theory. In Lyons, J., (Ed): *New Horizons in Linguistics*. Harmondsworth: Penguin.

Cobb, T. (2003) Analyzing Late Interlanguage with Learner Corpora: Quebec Replications of three European Studies. *The Canadian Modern Language Review,* 5, pp. 393–423.

Condon, N., & Kelly, P. (2002) Does Cognitive Linguistics Have Anything to Offer English Language Learners in their Efforts to Master Phrasal Verbs? *ITL Review of Applied Linguistics,* 133-134, pp. 205–231.

Conklin, K. and Schmitt, N. (2008) Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*. 29 (1), pp. 72-89.

Connor-Linton, J. (1995) Looking Behind the Curtain: What do L2 Composition Ratings Really Mean? *TESOL Quarterly*. 29 (4), pp. 762-765.

Cook, G. (1998) The Uses of Reality: A Reply to Ronald Carter. *ELT Journal*. 52 (1), pp. 57-63.

Cook, V. (2008) (4th Ed) *Second Language Learning and Language Teaching.* London: Edward Arnold.

Cooper, T. (1999) Processing of Idioms by L2 Learners of English. *TESOL Quarterly*. 33 (2), pp. 233-262.

Cowie, A. P. (1994). Phraseology. *The Encyclopedia of language and Linguistics*, *6*, pp. 3168-3171.

Cowie, A. P. (Ed.). (1998). *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Coxhead, A. (1998, 2000) A New Academic Word List. *TESOL Quarterly*. 34 (2), pp. 213-238.

Coxhead, A. (2011) The Academic Word List 10 Years on: Research and Teaching Implications. *TESOL Quarterly*. 45 (2), pp. 355-362.

Coxhead, A. and Byrd, P. (2007) Preparing Writing Teachers to Teach Vocabulary and Grammar of Academic Prose. *Journal of Second Language Writing* 16, pp. 129-147.

Cronbach, L.J. (1942) An Analysis of Techniques for Diagnostic Vocabulary Testing. *Journal of Educational Research,* 36, pp. 206-217.

Crossley, S.A., Salsbury, T., McNamara, D.S. and Jarvis, S. (2010; 2011a) Predicting Lexical Proficiency in Language Learner Texts Using Computational Indices. *Language Testing*. 28 (4), pp. 561-580.

Crossley, S.A., Salsbury, T., McNamara, D.S. and Jarvis, S. (2011b) What Is Lexical Proficiency? Some Answers from Computational Models of Speech Data. *TESOL Quarterly*, pp. 182-193.

Daller, H., Milton J. and Treffers-Daller, J. (eds.) (2007) Modelling and Assessing Vocabulary Knowledge. Cambridge: Cambridge University Press.

Daller, H. And Phelan, D. (2007) What is in a Teacher's Mind? Teacher Ratings of EFL Essays and Different Aspects of Lexical Richness. In Daller, H., Milton J. and J. Treffers-Daller (eds.) *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press, pp. 234-244.

Daller, H., Van Hout, R. and Treffers-Daller, J. (2003) Lexical Richness in the Spontaneous Speech of Bilinguals. *Applied Linguistics*. 24 (2), pp. 197-222.

Daller H. and Xue, H. (2007) Lexical Richness and the Oral Proficiency of Chinese EFL Students. In Daller, H., Milton J. and J. Treffers-Daller (eds.): *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

Darwin, C. M., & Gray, L. S. (1999) Going After the Phrasal Verb: An Alternative Approach to Classification. *TESOL Quarterly*. 33, pp. 65–83.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999) Studies in Language Testing (7) *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Dechert, H. W. (1983). How a story is done in a second language. *Strategies in interlanguage communication*, pp. 175-195.

Deese, J. (1965) The Structure of Associations in Language and Thought. Baltimore: Johns Hopkins University Press.

Demetriou, T. (2004) *Predicting Teachers' Ratings on the Basis of Different Measures of Lexical Richness and Error Analysis.* Linguistics Final Year Project, BA, University of the West of England, UK.

DeRocher, J.E. (1973) *The Counting of Words: A Review of the History, Techniques and Theory of Word Counts with Annotated Bibliography.* New York: Syracuse University Research Corp.

Douglas, D. (1994) Quantity and Quality in Speaking Test Performance. *Language Testing.* 11(2), pp. 125–44.

Douglas, D. and L. Selinker. (1992) Analysing Oral Proficiency Test Performance in General and Specific Purpose Contexts. *System.* 20, pp. 317–28.

Douglas, D. and L. Selinker. (1993) Performance on a General Versus a Field-specific Test of Speaking Proficiency by International Teaching Assistants. In D. Douglas and C. Chapelle (Eds): *A new Decade of Language Testing Research.* Alexandria, VA: TESOL Publications, pp. 235–56.

Dugast, D. (1979) *Vocabulaire et stylistique. I Théâtre et dialogue. Travaux de linguistique quantitative*. Geneva: Slatkine-Champion.

Duran, P., Malvern, D., Richards, B. and Chipere, N. (2004) Developmental Trends in Lexical Diversity. *Applied Linguistics*. 25 (2): 220-242.

Edwards, R. and Collins, L. (2011) Lexical Frequency Profiles and Zipf's Law. *Language Learning*. 61 (1), pp. 1-30.

Ehsanzadeh, S.J. (2012) Depth Versus Breadth of Lexical Repertoire: Assessing Their Roles in EFL Students' Incidental Vocabulary Acquisition. *TESL Canada.* 29 (2), pp. 24-41.

Ellis, N. C. (1996). Analyzing language sequence in the sequence of language acquisition: Some comments on Major and Ioup. *Studies in Second Language Acquisition. 18* (03), pp. 361-368.

Ellis, C.N., Simpson-Vlach, R. and Maynard, C. (2008) Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly*. 42 (3), pp. 375-396.

Engber, C.A. (1995) The Relationship of Lexical Proficiency to the Quality of ESL Compositions. *Journal of Second Language Writing*. 4 (2), pp. 139-155.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL-International Review of Applied Linguistics in Language Teaching. 6* (1-4), pp. 213-232.

Erling, E.J. and Richardson, J.T.E. (2010) Measuring the Academic Skills of University Students: Evaluation of a Diagnostic Procedure. *Assessing Writing*. 15 (3), pp. 177-193.

Erman B. and Warren, B. (2000) The Idiom Principle and the Open-Choice Principle. *Text*. 20, pp. 29–62.

Fan, M. (1991*) A Study of the Company Kept by a Selection of English Delexical Verbs and the Implications for Teaching of English in Hong Kong*. PhD, University of Durham, UK.

Fan, M. (2000) How Big is the Gap and How to Narrow it? An Investigation Into the Active and Passive Vocabulary knowledge of L2 Learners. *RELC Journal*. 31, pp. 105-119.

Fan, M. (2009) An Explanatory Study of Collocational Use by ESL Students- A Task Based Approach. *System.* 37 (1), pp. 110-123.

Farghal, M.and Obiedat, H. (1995) Collocations: A Neglected Variable in EFL. *International Review of Applied Linguistics in Language Teaching*. 33 (4), pp. 315–331.

Fatahipour, M. (2012) Exploring the Relationship Between Lexical Richness Measures and Standardised Vocabulary Knowledge Testing (PhD, University of the West of England, UK).

Fernando, C. (1996) *Idioms and Idiomaticity*. Oxford: Oxford University Press.

Field, A. (2005). *Discovering statistics with SPSS*. London: Sage.

Firth, J. R. (1952). Linguistic analysis as a study of meaning. *Selected papers of JR Firth*, *1959*. Cambridge: Cambridge University Press, pp. 12-26.

Fitzpatrick, T. and Clenton, J. (2010) The Challenge of Validation: Assessing the Performance of a Test of Productive Vocabulary. *Language Testing*. 27 (4), pp. 537–554.

Foster, P. (2001). Rules and Routines: A Consideration of their Role in the Task-based Language Production of Native and Non-native Speakers. In M.Bygate, P. Skehan & M.Swain (Eds.) *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing.* London, New York: Longman, pp. 75-94.

Frederiksen, J.R. (1982) A Componential Theory of Reading Skills and their Interactions. In Mislevy R.J. (Ed.): *Advances in the Psychology of Human Intelligence*, Vol. 1. Hillsdale, NJ: Lawrence Erlbaum.

Fries, C.C. and Traver, A.A. (1960) *English Word Lists.* Ann Arbor: George Wahr.

Gardner, D. (2007) Validating the Construct of Word in Applied Corpus-Based Vocabulary Research: A Critical Survey. *Applied Linguistics*. 28 (2), pp. 241–265

Gardner, D. and Davies, M. (2007) Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis. *TESOL Quarterly*. 41 (2), pp. 339-359.

George, D., & Mallery, P. (2003) SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 update (4th Ed.). Boston: Allyn & Bacon.

Goulden, NR. (1994) Relationship of Analytic and Holistic Methods to Raters' Scores for Speeches. *The Journal of Research and Development in Education, 27*: pp. 73–82.

Grabe, W. (1991): Current Developments in Second Language Reading Research. *TESOL Quarterly* 25,   pp. 375–406.

Grant, L. and Bauer, L. (2004) Criteria for Re-defining Idioms: Are we Barking up the Wrong Tree? *Applied Linguistics*. 25 (1), pp. 38-61.

Granger, S. (1998) *Learner English on Computer*. Longman: London.

Green, A. (2005) EAP Study Recommendations and Score Gains on the IELTS Academic Writing Test. *Assessing Writing*. 10 (1), pp. 44-60.

Greenbaum, S. 1974. Some verb-intensifier collocations in American and British English. *American Speech* 49, pp. 79–89.

Guiraud, P. (1960) *Problemes et Methods de la Statistique Linguistique*. Dordrecht: D. Reidel.

Halliday, M.A.K. (1966) Lexis as a Linguistic level. In: Bazell, C.E., Catford, J.C., Halliday, M.A.K., Robins, R.H. (Eds.): *In Memory of J.R. Firth.* London: Longmans, pp. 148–162.

Hamp-Lyons, L. (1995) Rating Nonnative Writing: The Trouble with Holistic Scoring. *TESOL Quarterly*. 29 (4), pp. 759-762.

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* 2: pp. 237–58.

Hatch, E. & Brown, C. (1995) *Vocabulary, Semantics and Language Education.* Cambridge. Cambridge University Press.

Hawkey, R. and Barker, F. (2004) Developing a Common Scale for the Assessment of Writing. *Assessing Writing*. 9 (2), pp. 122-159.

Heaps, H.S. (1978) *Information Retrieval: Computational and Theoretical Aspects.* New York: Academic Press.

Hellman, A.B. (2011) Vocabulary Size and Depth of Word Knowledge in Adult-onset Second Language Acquisition. *International Journal of Applied Linguistics*. 21 (2), pp. 162-182.

Henriksen, B. (1999). Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition*, *21* (2), pp. 303-317.

Herdan, G. (1960) *Type-token Mathematics: A Textbook of Mathematical Linguistics.* The Hague: Mouton De Gruyter.

Higgs, T. and Clifford, R. (1982) The Push Towards Communication. In T. V. Higgs (Ed): *Curriculum, Competence, and the Foreign Language Teacher*. Lincolnwood, IL: National Textbook Company, pp. 57–79.

Hill, J. (1999) *Collocational competence*. ETP (English Teaching Professional), Issue 11.

Hill, J. (2000). Revising Priorities: From Grammatical Failure to Collocational Success. In M. Lewis (Ed.): *Teaching Collocation: Further Developments in the Lexical Approach*. Hove, England: Language Teaching Publications, pp. 47–69.

Hirsch, D. and Nation, P. (1992) What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure? *Reading in a Foreign Language* 8(2), pp. 689-696.

Hoare, R. (2000) *Measuring Lexical Diversity: the TTR and D* (Dissertation, MA, University of the West of England, UK).

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Holmes, D.I. and **Singh, S. (**1996) A Stylometric Analysis of Conversational Speech of Aphasic Patients. *Literary and Linguistic Computing.* 11, pp. 133–40.

Howarth, P. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making* (Vol. 75). Tübingen: Max Niemeyer Verlag.

Howarth, P. (1998) The Phraseology of Learners' Academic Writing. In A. Cowie (ed.): *Phraseology: Theory, Analysis and Applications.* Oxford: Oxford University Press, pp. 161–86.

Howeler, M. (1972) Diversity of Word Usage as a Stress Indicator in an Interview Situation.

Journal of Psycholinguistic Research. 1, pp. 243–48.

Hsu, J. (2007) Lexical Collocations and Their Relation to the Online Writing of Taiwanese College English Majors and Non-English Majors. *Electronic Journal of Foreign Language Teaching*. 4 (2), pp. 192-209.

Hsu, J. Y., & Chiu, C. Y. (2008) Lexical Collocations and their Relation to Speaking Proficiency of College EFL learners in Taiwan. *Asian EFL Journal*. 10 (1), pp.181-204.

Hu, H.C.M & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language. 13* (1), pp. 403-30.

Hunston, S. and Francis, G. (2000) *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hwang, K. (1989). *Reading newspapers for the improvement of vocabulary and reading skills.* Thesis, MA, Victoria University of Wellington, New Zealand.

Hyland, K. (2008). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes.* 27 (1), pp. 4–21.

Hyland, K. and Tse, P. (2007) Is There an 'Academic Vocabulary'? *TESOL Quarterly*. 41 (2), pp. 235-253.

Iwashita, N., Brown, A., McNamara, T. and O'Hagan, S. (2008) Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*. 29 (1), pp. 24–49.

Jarvis, S. (2002) Short Texts, Best-Fitting Curves and New Measures of Lexical Diversity. *Language Testing.* 19 (1), pp. 57-84.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen.

Jiang, N. and Nekrasova, T.M. (2007) The Processing of Formulaic Sequences by Second Language Speakers. *The Modern Language Journal*. 91 (3), pp. 433-445.

Johnson, W. (1944) Studies in Language Behavior: A Program of Research. *Psychological Monographs*. 56, pp. 1-15.

Kaszubski, P. (2000) Selected *Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-Based Approach*. <http://main.amu.edu.pl/przemka/research.html>.

Kelly, P. (1989) Basic Components of Successful Foreign Language Vocabulary Learning. Paper presented at an international symposium on Vocabulary and Applied Linguistics. Universite Lumiere. Lyons 2.

Kennedy, G. (2003) Amplifier Collocations in the British National Corpus: Implications for English Language Teaching. *TESOL Quarterly*. 37 (3), pp. 467-487.

Kharma, N. and Hajjaj, A. (1997) *Errors in English among Arabic Speakers: Analysis and Remedy.* Beirut: York Press.

Kiss, G.R., Armstrong, C., Milroy, R. and Piper, J. (1973) An Associative Thesaurus of English and its Computer Analysis. In: A.J. Aitken, R.W. Bailey and N. Hamilton-Smith (Eds): *The Computer and Literary Studies*. Edinburgh: University Press.

Klee, T. (1992) Developmental and Diagnostic Characteristics of Quantitative Measures of Children's Language Production. *Topics in Language Disorders.* 12, pp. 28-41.

Kline, P. (1999) *The Handbook of Psychological Testing* (2nd Ed.). London: Routledge.

Knoch, U. (2009) Diagnostic Assessment of Writing: A Comparison of Two Rating Scales. *Language Testing.* 26 (2), pp. 275-304.

Knowles, G., & Don, Z. M. (2004). The notion of a 'lemma': Headwords, roots and lexical sets. *International Journal of Corpus Linguistics. 9* (1), pp. 69-81.

Kovecses, Z. and Szabo, P. (1996) Idioms: A View from Cognitive Semantics. *Applied Linguistics.* 17 (3), pp. 326-355.

Krashen, S. (2011) Academic Proficiency (Language and Content) and the Role of Strategies. *TESOL Journal.* 2 (4), pp. 381-393.

Laufer, B. (1992) Reading in a Foreign Language: How Does L2 Lexical Knowledge Interact with the Reader's General Academic Ability? *Journal of Research in Reading.* 15 (2), pp. 95-103.

Laufer, B. (1994) The Lexical Profile of Second Language Writing: Does it change over time? *RELC Journal.* 25, pp. 21-33.

Laufer, B. and Nation, P. (1995) Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics.* 16 (3), pp. 307-322.

Laufer, B. (1997) What's in a Word that Makes it Hard or Easy: Some Intralexical Factors that Affect the Learning of Words. In N. Schmitt & M. McCarthy (Eds.): *Vocabulary- Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Laufer, B. (1998) The Development of Passive and Active Vocabulary: Same or Different? *Applied Linguistics.* 19, pp. 255-271.

Laufer, B. (2005) Lexical Frequency Profiles: From Monte Carlo to the Real World (A Response to Meara 2005). *Applied Linguistics.* 26 (4), pp.582-588.

Laufer, B. and Nation, P. (1999) A Vocabulary-Size Test of Controlled Productive Ability. *Language Testing.* 16 (1), pp. 33-51.

Laufer, B., Elder, C., Hill, K. and Congdon, P. (2004) Size and Strength: Do we Need Both to Measure Vocabulary Knowledge? *Language Testing.* 21 (2), pp. 202-226.

Laufer, B. and Waldman, T. (2011) Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning.* 61 (2), pp. 647-672.

Lee, Y.W., Gentile, C. and Kantor, R. (2009) Toward Automated Multi-trait Scoring of Essays: Investigating Links Among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics.* 31 (3), pp. 391–417.

Leech, G. (2001). The role of frequency in ELT: New corpus evidence brings a re-appraisal. *Foreign Language Teaching and Research. 33*(5), pp. 328-339.

Lennon, P. (1991) Error: Some Problems of Definition, Identification, and Distinction. *Applied Linguistics*. 12 (2), pp. 180-196.

Lewis, M. (1993) *The Lexical Approach*. Hove, England: Language Teaching Publications.

Lewis, M. (2000) *Teaching Collocation: Further Developments in the Lexical Approach*. London: Language Teaching Publications.

Lewis, M. (2002). *Implementing the Lexical Approach*. Boston, MA: Thomson/Heinle.

Lewis, M. (2008) The Idiom Principle in L2 English: Assessing Elusive Formulaic Sequences as Indicators of Idiomaticity, Fluency and Proficiency. Thesis, PhD, Stockholm University.

Li J. and Schmitt, N. (2009) The Acquisition of Lexical Phrases in Academic Writing: A Longitudinal Case Study. *Journal of Second Language Writing*. 18 (2), pp. 85-102.

Liu, D. (2003) The Most Frequently Used Spoken American English Idioms: A Corpus Analysis and Its Implications. *TESOL Quarterly*. 37 (4), pp. 671-700.

Liu, D. (2010). Going Beyond Patterns: Involving Cognitive Analysis in the Learning of Collocations. *TESOL Quarterly*. 44 (1), pp. 4-30.

Liu, D. (2011). The Most Frequently Used English Phrasal Verbs in American and British English: A Multicorpus Examination. *TESOL Quarterly*. 45 (4), pp. 661-688.

Lorenz, T.R. (1999) *Adjective Intensification – Learners Versus Native Speakers. A Corpus Study of Argumentative Writing*. Amsterdam and Atlanta: Rodopi.

Lorenzo Dus, N. (2007) Best of Both Worlds? Combined Methodological Approaches to the Assessment of Vocabulary in Oral Proficiency Interviews (220- 233). In Daller, H., Milton J. and J. Treffers-Daller (Eds.): *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Maas, H. D. (1972) Zusammenhang Zwischen Wortschatzumfang und Länge Eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*. 8: 73-79.

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Magnan, S. (1988) Grammar and the ACTFL Oral Proficiency Interview: Discussion and Data. *The Modern Language Journal*. 72 (3), pp. 266–76.

Malvern, D. and Richards, B. (1997) A New Measure of Lexical Diversity. In: Ryan, A. and Wray, A. (Eds): *Evolving Models of Language. Papers from the Annual Meeting of BAAL held at the University of Swansea, Wales, September 1996,* 58-71. Clevedon: Multilingual Matters.

Malvern, D. and Richards, B. (2002) Investigating Accommodation in Language Proficiency Interviews Using a New Measure of Lexical Diversity. *Language Testing.* 19 (1), pp. 85-104.

Malvern D., B. Richards, N. Chipere and Duran, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. New York: Palgrave Macmillan.

Martinez, R. and A. Murphy, V. (2011) Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *TESOL Quarterly*. 45 (2), pp. 267-290.

Martinez, R., & Schmitt, N. (2010). Invited commentary: vocabulary. *Language Learning & Technology*. *14* (2), pp. 26-29.

Martinez, R. and Schmitt, N. (2012) A Phrasal Expressions List. *Applied Linguistics*. 33 (3), pp. 1-23.

Mayor, B.Hewings, A, North, S, Swann, J and Coffin, C. (2002) *A Linguistic Analysis of Chinese and Greek L1 Scripts for IELTS Academic Writing Task 2.* IELTS British Council Research Programme.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. and O'Dell, F. (2005) *English Collocations in Use*. Cambridge: Cambridge University Press.

McCarthy, P.M and Jarvis, S. (2007) *Vocd*: A Theoretical and Empirical Evaluation. *Language Testing*. 24 (4), pp. 459–488.

McCarthy, P. & Jarvis, S. (2010) MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*. 42, pp. 381-392.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

McGovern, P. and Walsh, S. (2006). *IELTS Research Reports Volume 6.* Australia: IELTS Australia Pty Limited.

McGovern, P. and Walsh, S. (2007) *IELTS Research Reports Volume 7.* Australia: IELTS Australia Pty Limited.

McNamara, D.S., Crossley, S.A., and McCarthy, P.M. (2010). The Linguistic Features of Quality Writing. *Written Communication. 27* (3-4), pp. 221-246.

McNamara, T. (2011) Applied Linguistics and Measurement: A Dialogue. *Language Testing*. 28 (4), pp. 435-440.

McKee, G., D.D. Malvern and Richards, B.J. (2000) Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing*. 15, pp. 323-38.

Meara, P. (1980). Vocabulary Acquisition: A Neglected Aspect of Language Learning. *Language Teaching*. 13, pp. 221-246.

Meara, P. (1983). Word associations in a foreign language. *Nottingham Linguistics Circular*. *11* (2), pp. 29-38.

Meara, P. (1990) A Note on Passive Vocabulary. *Second Language Research*. 6, pp.150-154.

Meara, P. (1996) The Dimensions of Lexical Competence'. In G. Brown, K. Malmkjaer, and J. Williams (Eds.): *Performance and Competence in Second Language Acquisition.* Cambridge: Cambridge University Press, pp. 35-53.

Meara, P. (2005) Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics*. 26 (1), pp. 32-47.

Meara, P. and Bell, H. (2001) P_Lex: A Simple Effective Way of Describing the Lexical Characteristics of Short L2 Texts. *Prospect.* 16 (3), pp. 5-17.

Meara, P. and Buxton, B. (1987) An Alternative to Multiple Choice Vocabulary Tests. *Language Testing*. 4 (2), pp. 142-154.

Meara, P. And Fitzpatrick, T. (2000) Lex30: An Improved Method of Assessing Productive Vocabulary in an L2. *System.* 28 (1), pp. 19-30.

Melka, F. (1997) Receptive vs. Productive Aspects of Vocabulary. In Schmitt, N. And McCarthy, M. (Eds): *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: Cambridge University Press, pp. 84-102.

Michéa, R. (1969) Répétition et variété dans l'emploi des mots. *Bulletin de la Société de Linguistique de Paris*, 1–24.

Mickan, P. (2003) What is Your Score? An Investigation into Language Descriptors from Rating Written Performance. *IELTS Research Reports* 5 (3). Canberra, Australia: IDP IELTS Australia.

Mickan, P. and Slater, S. (2003) Text Analysis and the Assessment of Academic writing. *IELTS Research Reports* 4 (2), pp. 59-88.

Millar, N. (2011).The Processing of Malformed Formulaic Language. *Applied Linguistics*. 32 (2), pp. 129-148.

Milton, J. (2009*) Measuring Second Language Vocabulary Acquisition*. Bristol, England: Multilingual Matters.

Möbarg, M. (1997) Acquiring, Teaching and Testing Vocabulary. *International Journal of Applied Linguistics*. 7 (2), pp. 201-222.

Mochida, K. and Harrington, M. (2006) The Yes/No Test as a Measure of Receptive Vocabulary Knowledge. *Language Testing*. 23 (1), pp. 73–98.

Moon, R. (1997) Vocabulary Connections: Multi-word Items in English. In N. Schmitt & M. McCarthy (Eds.): *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press, pp. 40-63.

Moon, R. (1998) *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Moore, T. and Morton, J. (2005) Dimensions of Difference: A Comparison of University Writing and IELTS Writing. *Journal of English for Academic Purposes*. 4 (1), pp. 43-66.

Morris, L. and Cobb, T. (2004) Vocabulary Profiles as Predictors of the Academic Performance of Teaching English as a Second Language Trainees. *System*. 32 (1), pp. 75-87.

Morris, L., and Tremblay, M. (2002) The Interaction of Morphology and Lexis in the Development in ESL Learners. Paper given at the Canadian Association of Applied Linguistics Conference, Toronto.

Nation, P. (l983) Testing and Teaching Vocabulary. *Guidelines.* 5 (1), pp.12–25.

Nation, P. (1990) *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, P. (2001) *Learning Vocabulary in Another language*. Cambridge: Cambridge University Press.

Nation, P. (2008) Vocabulary. In D.Nunan (Ed): *Practical English Language Teaching.* New York: McGraw Hill, pp. 129-152.

Nation, I. S. P., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher*. *31* (7), pp. 9-13.

Nation, P. and R. Waring (1997) Vocabulary Size, Text Coverage and Word Lists. In Schmitt, N. and M. McCarthy (Eds.): *Vocabulary: Description, Acquisition and Pedagogy*: Cambridge, Cambridge University Press, pp. 6-19.

Nattinger J.R and DeCarrico, J.S. (1992) *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nekrasova, T.M. (2009) English L1 and L2 Speakers' Knowledge of Lexical Bundles. *Language Learning*. 59 (3), pp. 647-686.

Nesselhauf, N. (2003) The Use of Collocations by Advanced Learners of English and some implications for teaching. *Applied Linguistics*. 24 (2), pp. 223-242.

Nesselhauf, N. (2004). Learner Corpora and their Potential for Language Teaching. *How to use corpora in language teaching*, *12*.

Nesselhauf, N. (2005) *Collocations in a Learner Corpus*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nunnally, J.C., Durham, R.L., Lemond, L.C.and Wilson, W.H. (1975). *Introduction to Statistics for Psychology and Education.* McGraw-Hill Book.

Ohlrogge, A. (2009) Formulaic Expressions in Intermediate EFL Writing Assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (Eds): *Formulaic Language Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 375–86.

Official IELTS Practice Materials 2003

O'Loughlin, K. (2001) *The equivalence of semi-direct speaking tests. Cambridge.* University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Owen, A.J. and **Leonard, L.B.** (2002) Lexical Diversity in the Spontaneous Speech of Children with Specific Language Impairment: Application of *D. Journal of Speech and Hearing Research.* 45, pp. 927–37.

Palmer, H. E. (1933). *Second Interim Report on English Collocations*. Tokyo, Japan: Kaitakusha.

Paribakht, T.S., & Wesche, M. (1993) Reading Comprehension and Second Language Development in a Comprehension-Based ESL Program. *TESL Canada Journal*. 11 (1), pp. 9 – 29.

Paribakht, T.S. and **Wesche, M.** (1997) Vocabulary Enhancement Activities and Reading for Meaning in Second Language Vocabulary Acquisition. In Coady, J. and Huckin, T., (Eds): *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press, pp. 174–200.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching* (Vol. 2). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pawley A. and Syder, F.H. (1983) Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In: J.C. Richards and R.W. Schmidt (Eds): *Language and Communication*. London: Longman, pp. 191–225.

Pearson, D., Hiebert, E.H. and Kamil, M.L. (2007) Theory and Research Into Practice: Vocabulary Assessment: What we Know and What we Need to Learn. *Reading Research Quarterly*. 42 (2), pp. 282-296.

Pike, L.W. (1979) An Evaluation of Alternative Item Formats for Testing English as a Foreign Language. *TOEFL Research Reports, 2.* Princeton, NJ: Educational Testing Service.

Postman, L. and Keppel, G. (1970) Norms of word associations. New York: Academic Press.

Qian, D. (1999) Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *The Canadian Modern Language Review.* 56, pp. 283-307

Qian, D. and Schedl, M. (2004) Evaluation of an In-Depth Vocabulary Knowledge Measure for Assessing Reading Performance. *Language Testing.* 21 (1), pp. 28-52.

Read, J. (1989) *Towards a Deeper Assessment of Vocabulary Knowledge*. Paper presented at the 8th World Congress of Applied Linguistics. Sydney, NSW, Australia, August, 1987. ED 301048. Washington, DC: ERIC Clearinghouse on Languages and Linguistics

Read, J. (1993) The Development of a New Measure of L2 Vocabulary Knowledge. *Language Testing.* 10 (3), pp. 355-371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. *Validation in language assessment.* pp. 41-60.

Read, J. (2000) *Assessing Vocabulary (edn).* Cambridge: Cambridge University Press

Read, J. (2004) Plumbing the Depths: How Should the Construct of Vocabulary Knowledge be Defined. In: P. Bogaards and B. Laufer, Editors, Vocabulary in a Second Language. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 209–227.

Read, J. and Chapelle, C.A. (2001) A Framework for Second Language Vocabulary Assessment. *Language Testing.* 18 (1), pp. 1-32.

Read, J. and P. Nation (2002) An Investigation of the Lexical Dimension of the IELTS Speaking Test. *IELTS Research Reports 6:* 207-231

Revelle, W., and Zinbarg, R. (2009) Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika.* 74 (1), pp. 145–154.

Richards, J. (1974). *Error analysis: Perspectives on second language acquisition*. Longman Pub Group.

Richards, J. (1976) The Role of Vocabulary teaching. *TESOL Quarterly.* 10, pp. 77-89

Ruegg, R., Fritz, E. and Holland, J. (2011) Rater Sensitivity to Qualities of Lexis in Writing. *TESOL Quarterly.* 45 (1), pp. 63-80.

Schmitt, N. (1994) Vocabulary Testing: Questions for Test Development with Six Examples of Tests of Vocabulary Size and Depth. *Thai TESOL Bulletin.* 6 (2), pp. 9-16

Schmitt, N. (1995) A Fresh Approach to Vocabulary Using a Word Knowledge Framework. *RELC Journal.* 26, pp. 86–94.

Schmitt, N. (1998a) Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning.* 48, pp. 281–317.

Schmitt, N. (1999) The Relationship Between TOEFL Vocabulary Items and Meaning, Association, Collocation and Word-class Knowledge. *Language Testing.* 16 (2): 189–216.

Schmitt, N. (Ed.) (2004) *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Schmitt, N. (2010) *Researching Vocabulary: A Vocabulary Research Manual.* Basingstoke: Palgrave Macmillan.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. *Formulaic sequences: Acquisition, processing and use*, 1-22.

Schmitt, N., Jiang, X. and Grabe, W. (2011) The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal.* 95 (1), pp. 26-43.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in second language acquisition*, *19* (01), pp. 17-36.

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), pp.55-88.

Schmitt, N., Wun Ching Ng, J. and Garras, J. (2011) The Word Associates Format: Validation Evidence. *Language Testing.* 28 (1), pp. 105–126.

Schmitt, N. and Zimmerman, C.B. (2002) Derivative Word Forms: What Do Learners Know? *TESOL Quarterly.* 36 (2), pp. 145-171.

Schoonen, R. (2001) Book Review: Assessing Vocabulary. *Language Testing.* 18 (1), pp. 118-125.

Schoonen, R. (2005) Generalizability of Writing Scores: An Application of Structural Equation Modeling. *Language Testing.* 22 (1), pp. 1-30.

Schoonen, R. and Verhallen, M. (2008) The Assessment of Deep Word Knowledge in Young First and Second Language learners. *Language Testing.* 25 (2): 211-237.

Scott, M., (1996) *WordSmith Tools*. Oxford: Oxford University Press.

Shen, Z. (2008) The Roles of Depth and Breadth of Vocabulary Knowledge in EFL Reading Performance. *Asian Social Science.* 4 (12), pp. 135

Shin, D. (2007) The High Frequency Collocations of Spoken and Written English. *English Teaching.* 62 (1), pp.199–218.

Shin, D. and Nation, P. (2008) Beyond Single Words: The Most Frequent Collocations in Spoken English. *ELT Journal.* 62 (4), pp. 339-348.

Sichel, H.S. 1975: On a distributive law for word frequencies. *Journal of the American Statistical Association.* 70, pp. 542–47.

Silverman, S. and **Bernstein Ratner, N.** (2000) Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist.* 11, pp. 45–72.

Simpson, R. and Mendis, D. (2003) A Corpus-Based Study of Idioms in Academic Speech. *TESOL Quarterly*. 37 (3), pp. 419-441.

Simpson-Vlach, R. and C. Ellis, N. (2010) An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*. 31 (4), pp. 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. and Renouf, A. (1988) A Lexical Syllabus for Language Learning. In R.Carter & M.McCarthy (Eds): *Vocabulary and Language Teaching*. London, New York: Longman, pp. 140-160.

Singleton, D. M. (1999). *Exploring The Second Language Mental Lexicon*. Ernst Klett Sprachen.

Siyanova, A. and Schmitt, N. (2008) L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *The Canadian Modern Language Review*. 64 (3), pp. 429-458.

Smadja, F., and McKeown, K. (1991) Using Collocations for Language Generation. *Computational Intelligence*. 7, pp. 229–239.

Somers, H.H. (1966): Statistical methods in literary analysis. In Leeds, J., (Ed): *The computer and literary style*. Kent, OH: Kent State University, pp. 128–40.

Sorhus, H.B. (1977) To Hear Ourselves – Implications for Teaching English as a Second Language. *English Language Teaching Journal* .31 (3), pp. 211–221.

Stahl, S. (1999) *Vocabulary Development.* Boston: Brookline Books

Stengers, I. (2009). *Au temps des catastrophes: résister à la barbarie qui vient*. La Découverte.

Thompson, G. & **Thompson, J**. (1915) Outlines of a Method for the Quantitative Analysis of Writing Vocabularies. *British Journal of Psychology. 8, pp.* 52-69

Thordardottir, E.T., and Ellis Weismer, S. (2001) High Frequency Verbs and Verb Diversity in the Spontaneous Speech of School-Age Children with Specific Language Impairment. *International Journal of Language and Communication Disorders*. 36: 221-244.

Tidball, F.T and Treffers –Daller, J. (2007) Exploring Measures of Vocabulary Richness in Semi Spontaneous French Speech: A Quest for the Holy Grail? In Daller, H., Milton J. and J. Treffers-Daller (Eds.): *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP.

Treffers-Daller, J. (2013) Measuring Lexical Diversity Among L2 Learners of French: An Exploration of the Validity of D, MTLD and HD-D as Measures of Language Ability. In: Jarvis, S. and Daller, M. (Eds.): *Vocabulary Knowledge: Human Ratings and Automated Measures.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Treffers-Daller, J., Daller, H. M., Malvern, D., Richards, B., Meara, P., & Milton, J. (2008). Introduction: Special issue on knowledge and use of the lexicon in French as a second language. *Journal of French Language Studies*. *18* (03), pp. 269-276.

Tremblay, A., Derwing, B., Libben, G. and Westbury, C. (2011) Processing Advantages of Lexical Bundles: Evidence from Self-Paced Reading and Sentence Recall Tasks. *Language Learning*. 61 (2), pp. 569-613.

Turlik, J. (2008) *Identifying Change in Lexical Richness in Second Language English Academic Essay Written by Arabic First Language Students*. Thesis, PhD, University of the West of England, UK.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*. *32* (5), pp. 323-352.

Uysal, H. (2010) A Critical Review of the IELTS Writing Test. *ELT Journal*. 64 (3), pp. 314-320.

Van Hout, R. and Vermeer, A. (2007) Comparing Measures of Lexical Richness. In Daller, H., Milton J. and J. Treffers-Daller (Eds.): *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Van Lancker-Sidtis, D. and Rallon, G. (2004) Tracking the Incidence of Formulaic Expressions in Everyday Speech: Methods for Classification and Verification. *Language & Communication*. 24 (3), pp. 207-240.

Verhallen, M. and Schoonen, R. (1998) Lexical Knowledge in L1 and L2 of Third and Fifth Graders. *Applied Linguistics.* 19(4), pp. 452-470

Vermeer, A. (1992) Exploring the Second Language Learner Lexicon. In: Verhoeven, L. and DeJong, J.H.A.L., (Eds): *The Construct of Language Proficiency.* Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 147-162

Vermeer, A. (2000) Coming to Grips with Lexical Richness in Spontaneous Speech Data. *Language Testing*. 17 (1), pp. 65-83.

Vermeer, A. (2001) Breadth and Depth of Vocabulary in Relation to L1/L2 Acquisition and Frequency of Input. *Applied Psycholinguistics.* 22, pp. 217-234

Vogel Sosa A. and MacFarlane, J. (2002) Evidence for Frequency-Based Constituents in the Mental Lexicon: Collocations Involving the Word *of. Brain and Language*. 83 (2), pp. 227-236.

Walker, C.P. (2011) A Corpus-Based Study of the Linguistic Features and Processes Which Influence the Way Collocations Are Formed: Some Implications for the Learning of Collocations. *TESOL Quarterly*. 45 (2), pp. 291-312.

Waring, R. (1997) A Comparison of the Receptive and Productive Vocabulary Sizes of Some Second Language Learners. *Immaculata* (Notre Dame Seishin University, Okayama). 1, pp. 53-68.

Webb, S. (2005) Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition.* 27, pp.33-52

Webb, S. (2008) Receptive and Productive Vocabulary Sizes of L2 Learners. *Studies in Second Language Acquisition*. 30, pp. 79-95.

Webb, S. and Kagimoto, E. (2009) The Effects of Vocabulary Learning on Collocation and Meaning. *TESOL Quarterly*. 43 (1), pp. 55-77.

Webb, S. and Kagimoto, E. (2011) Learning Collocations: Do the Number of Collocates, Position of the Node Word, and Synonymy Affect Learning? *Applied Linguistics*. 32 (3), pp. pp. 259-276.

Weigle, S. C. (2002) *Assessing writing*. Cambridge: Cambridge University Press.

Wesche, M. and Paribakht, S.T. (1996) Assessing Second Language Vocabulary Knowledge: Depth versus Breadth. *The Canadian Modern Language Review*. 53(1), pp. 13-40

West, M. (1953) *A General Service List of English Words*. London: Longman, Green.

Widdowson, H. G. (2000) On the Limitations of Linguistics Applied'. *Applied Linguistics*. 21, pp. 3–25.

Wolter, B. (2002) Assessing Proficiency Through Word Associations: Is There Still Hope? *System.* 30 (3), pp. 315-329

Wolter, B. (2006). Lexical Network Structures and L2 Vocabulary Acquisition: The Role of L1 Lexical Conceptual knowledge. *Applied Linguistics*. *27* (4), pp. 741-747.

Wolter, B. and Gyllstad, H. (2011) Collocational Links in the L2 Mental Lexicon and the Influence of L1 Intralexical Knowledge. *Applied Linguistics*. 32 (4), pp. 430-449.

Wood, M. M. (1986). *A definition of idiom* (Vol. 324). Indiana University Linguistics Clubs.

Wouden, T. V. D. (1997) *Negative Contexts: Collocation, Polarity and Multiple Negation*. London, England: Routledge.

Wray, A. (1999) Formulaic Language in Learners and Native Speakers. *Language Teaching*. 32 (04), pp. 213-231.

Wray, A. (2000) Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics*. 21 (4), pp. 463-489.

Wray, A. and Perkins, M.R. (2000) The Functions of Formulaic Language: An Integrated Model. *Language & Communication.* 20 (1), pp. 1–28.

Wray, A. (2002) *Formulaic Language and the Lexicon.* United States of America, New York: Cambridge University Press

Wray, A. (2008) *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Xi, X. (2010) Automated Scoring and Feedback Systems: Where are we and Where are we Heading? *Language Testing*. 27 (3), pp. 291-300.

Xue, G., & **Nation,** I. S. P. (1984). A university word list. *Language learning and communication*. *3* (2), pp. 215-229.

Yamashita, J. and Jiang, N. (2010) L1 Influence on the Acquisition of L2 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *TESOL Quarterly*. 44 (4), pp. 647-668.

Yu, G. (2007) Lexical Diversity in MELAB Writing and Speaking Task Performances. *SPAAN Fellow Working Papers in Second or Foreign Language Assessment.* Volume 5 University of Michigan, English Language Institute.

Yu, G. (2009) Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*. 31 (2), pp. 236-259.

Yu, X. (2009) A Formal Criterion for Identifying Lexical Phrases: Implication from a Classroom Experiment. *System.* 37 (4), pp. 689-699

Yule, G.U. (1944) *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

Zareva, A. (2005) Models of Lexical Knowledge Assessment of Second Language Learners of English at Higher Levels of Language Proficiency. *System.* 33 (4), pp. 547-562.

Zughoul, M. And Osman Kambal, M. (1983) Objective Evaluation of EFL Composition. *International Review of Applied Linguistics in Language Teaching.* 21 (2), pp. 87-104

Zyzik, E. (2011) Second Language Idiom Learning: The Effects of Lexical Knowledge and Pedagogical Sequencing. *Language Teaching Research.* 15 (4), pp. 413-433.

Electronic References

IELTS Website http://www.ielts.org/PDF/113313_AC_sample_scripts.pdf   Date accessed: 20 July 2014

Bristol Centre for Linguistics: http://www1.uwe.ac.uk/cahe/research/bristolcentreforlinguistics/researchatbcl/iclru.aspx Date accessed: 15 July 2015

Websites

British Council Website: http://www.britishcouncil.org Date accessed: 15 June 2006

Cardiff University Website, http://www.cardiff.ac.uk   Date accessed: 22 September 2010

IELTS Help Now: www.ielts.org Date accessed: 17 January 2009

IELTS Band Descriptors: http://www.ieltshelpnow.com/ielts_grading.html Date accessed: 15 October 2007

Using English for Academic Purposes: A guide for Students in Higher Education- Vocabulary in EAP Gillet, A. http://www.uefap.co.uk   Date accessed: 05 November 2011

VocabProfile (http://www.lextutor.ca/vp/eng/).

Cobb, T. *Web VocabProfile* [accessed July 2009 from http://www.lextutor.ca/vp/], an adaptation of Heatley, Nation & Coxhead's (2002) *Range*.

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). RANGE and FREQUENCY programs. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

APPENDICES

Appendix 1

CONSENT FORM

I have read the participants' information sheet about Theodosia Demetriou's Project on the lexical richness of Greek-Cypriot EFL learners and teacher ratings in Cyprus and I am happy to participate in the project, as outlined in the participants' information sheet and for the data to be published in anonymous form.

Name:

Place:

Date:

Signature:

PARTICIPANT INFORMATION SHEET (STUDENTS)

Project on the lexical richness of Greek-Cypriot EFL learners and teacher ratings in Cyprus: how to predict IELTS Writing and Speaking scores using measures of lexical richness

You are invited to take part in a study on lexical richness and IELTS teacher ratings in Cyprus. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

What is the purpose of the study?

This is my MA project which is a follow-up study of a smaller scale one for my Final Year Linguistics project on Lexical richness and teacher ratings. The main aim of the study is to see which measure of lexical richness correlates higher with teacher ratings and find the most appropriate one to use for predicting IELTS writing and speaking exam scores.

Why have I been chosen?

I have been given the permission from your English teacher.

Do I have to take part?

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part

you are still free to withdraw at any time, or a decision not to take part, will not affect you in any way.

What will happen to me if I take part and what do I have to do?

You will be asked to write an essay from an IELTS past exam paper Writing Task. You will also have to conduct an interview exactly how it is done in IELTS Speaking test. You will be asked some general questions and then you will be given a piece of paper on a specific subject , have some time to prepare and think about it and will be asked to speak on that subject for approximately 2 minutes without any interruptions. There will be some follow up questions. In addition, there is a questionnaire to fill in. (for background information)

What are the possible disadvantages and risks of taking part?

I will give you an anonymous number that will be used instead your name on any of the information that you provide. Personal information that you will reveal will not be given to anyone else.

What are the possible benefits of taking part?

You will get some practice before your exam and you will be contributing to a research that may lead to a very useful diagnostic tool for teachers and students.

What if something goes wrong?

If you have any questions or complaints you can always contact me or my institution for further details.

Will my taking part in this study be kept confidential?

All information which is collected about you during the course of the research will be kept strictly confidential. Any information about you which leaves the workplace will have your name and address removed so that you cannot be recognized from it.

What will happen to the results of the research study?

I hope to write my thesis presenting the result s of the study and hopefully get it published. I can provide a summary of the results for all those who participated.

Who is organising and funding the research?

I am organizing this study for my MA dissertation. The research is self-funded at the moment but I am applying for funds from different funding bodies in the United Kingdom.

Contact for further information

Theodosia Demetriou

University of the West of England, Bristol

Faculty of Humanities, Languages and Social Sciences

School of Languages, Linguistics and Area Studies

Frenchay Campus

Coldharbour Lane

Bristol

BS16 1QY

Tel. 00447812760030

Email: Theodosia.Demetriou@uwe.ac.uk

QUESTIONNAIRE

Gender (Male/ Female):

Age:

What is your father's profession? (if retired state their profession before retirement)

What is your mother's profession?

How many years have you been learning English for:

Have you ever lived in an English speaking country?

Which school do you go to? (public or private school)

Do you use English with your friends?

1	**2**	**3**	**4**	**5**
Never	**Rarely**	**Occasionally**	**Frequently**	**Always**

Do you use English at home with your parents?

1	**2**	**3**	**4**	**5**
Never	**Rarely**	**Occasionally**	**Frequently**	**Always**

Appendix 2

IELTS Speaking band descriptors (public version)

BRITISH COUNCIL · IELTS AUSTRALIA · UNIVERSITY of CAMBRIDGE ESOL Examinations

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• is easy to understand throughout; with L1 accent having minimal effect on intelligibility • uses a wide range of phonological features to convey meaning effectively
7	• speaks at length without noticeable effort or loss of coherence • uses a range of connectives and discourse markers with some flexibility • may demonstrate language-related hesitation at times, or some repetition and/or self-correction	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures, though these rarely cause comprehension problems	• can be understood throughout, though mispronunciation may occasionally cause momentary strain for the listener

IELTS Speaking band descriptors (public version)

BRITISH COUNCIL   IELTS AUSTRALIA   UNIVERSITY of CAMBRIDGE ESOL Examinations

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
5	• usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• produces some acceptable features of English pronunciation but overall control is limited and there can be severe strain for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• speech is often unintelligible
1	• no communication possible • no rateable language			
0	• does not attend			

IELTS Writing Band Descriptors: Task 2 (Public Version)

Band	Task response	Coherence and cohesion	Lexical resource	Grammatical range and accuracy
9	• fully addresses all parts of the task • presents a fully developed position in answer to the question with relevant, fully extended and well supported ideas	• uses cohesion in such a way that it attracts no attention • skilfully manages paragraphing	• uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'	• uses a wide range of structures with full flexibility and accuracy; rare minor errors occur only as 'slips'
8	• sufficiently addresses all parts of the task • presents a well-developed response to the question with relevant, extended and supported ideas	• sequences information and ideas logically • manages all aspects of cohesion well • uses paragraphing sufficiently and appropriately	• uses a wide range of vocabulary fluently and flexibly to convey precise meanings • skilfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation • produces rare errors in spelling and/or word formation	• uses a wide range of structures • the majority of sentences are error-free • makes only very occasional errors or inappropriacies
7	• addresses all parts of the task • presents a clear position throughout the response • presents, extends and supports main ideas, but there may be a tendency to over-generalise and/or supporting ideas may lack focus	• logically organises information and ideas; there is clear progression throughout • uses a range of cohesive devices appropriately although there may be some under-/over-use • presents a clear central topic within each paragraph	• uses a sufficient range of vocabulary to allow some flexibility and precision • uses less common lexical items with some awareness of style and collocation • may produce occasional errors in word choice, spelling and/or word formation	• uses a variety of complex structures • produces frequent error-free sentences • has good control of grammar and punctuation but may make a few errors
6	• addresses all parts of the task although some parts may be more fully covered than others • presents a relevant position although the conclusions may become unclear or repetitive • presents relevant main ideas but some may be inadequately developed/unclear	• arranges information and ideas coherently and there is a clear overall progression • uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical • may not always use referencing clearly or appropriately • uses paragraphing, but not always logically	• uses an adequate range of vocabulary for the task • attempts to use less common vocabulary but with some inaccuracy • makes some errors in spelling and/or word formation, but they do not impede communication	• uses a mix of simple and complex sentence forms • makes some errors in grammar and punctuation but they rarely reduce communication
5	• addresses the task only partially; the format may be inappropriate in places • expresses a position but the development is not always clear and there may be no conclusions drawn • presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail	• presents information with some organisation but there may be a lack of overall progression • makes inadequate, inaccurate or over-use of cohesive devices • may be repetitive because of lack of referencing and substitution • may not write in paragraphs, or paragraphing may be inadequate	• uses a limited range of vocabulary, but this is minimally adequate for the task • may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader	• uses only a limited range of structures • attempts complex sentences but these tend to be less accurate than simple sentences • may make frequent grammatical errors and punctuation may be faulty; errors can cause some difficulty for the reader
4	• responds to the task only in a minimal way or the answer is tangential; the format may be inappropriate • presents a position but this is unclear • presents some main ideas but these are difficult to identify and may be repetitive, irrelevant or not well supported	• presents information and ideas but these are not arranged coherently and there is no clear progression in the response • uses some basic cohesive devices but these may be inaccurate or repetitive • may not write in paragraphs or their use may be confusing	• uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task • has limited control of word formation and/or spelling; errors may cause strain for the reader	• uses only a very limited range of structures with only rare use of subordinate clauses • some structures are accurate but errors predominate, and punctuation is often faulty
3	• does not adequately address any part of the task • does not express a clear position • presents few ideas, which are largely undeveloped or irrelevant	• does not organise ideas logically • may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas	• uses only a very limited range of words and expressions with very limited control of word formation and/or spelling • errors may severely distort the message	• attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
2	• barely responds to the task • does not express a position • may attempt to present one or two ideas but there is no development	• has very little control of organisational features	• uses an extremely limited range of vocabulary; essentially no control of word formation and/or spelling	• cannot use sentence forms except in memorised phrases
1	• answer is completely unrelated to the task	• fails to communicate any message	• can only use a few isolated words	• cannot use sentence forms at all
0	• does not attend • does not attempt the task in any way • writes a totally memorised response			

Appendix 3

Reliability Analysis for Written Overall

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	42	100.0
	Excluded[a]	0	.0
	Total	42	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.578	.584	2

Item Statistics

	Mean	Std. Deviation	N
Wr Overall EX1	5.1905	.52906	42
Wr Overall EX2	5.6429	.62748	42

Inter-Item Correlation Matrix

	Wr Overall EX1	Wr Overall EX2
Wr Overall EX1	1.000	.412
Wr Overall EX2	.412	1.000

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Wr Overall EX1	5.6429	.394	.412	.170	.[a]
Wr Overall EX2	5.1905	.280	.412	.170	.[a]

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
10.8333	.947	.97322	2

Appendix 4

Reliability Analysis for Oral Overall

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	42	100.0
	Excluded[a]	0	.0
	Total	42	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.795	.800	3

Item Statistics

	Mean	Std. Deviation	N
Or Overall EX1	5.1786	.57192	42
Or Overall EX2	6.1190	.77938	42
Or Overall EX3 (main)	5.2506	.64156	42

Inter-Item Correlation Matrix

	Or Overall EX1	Or Overall EX2	Or Overall EX3 (main)
Or Overall EX1	1.000	.608	.483
Or Overall EX2	.608	1.000	.621
Or Overall EX3 (main)	.483	.621	1.000

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Or Overall EX1	11.3696	1.640	.612	.388	.757
Or Overall EX2	10.4292	1.093	.714	.509	.649
Or Overall EX3 (main)	11.2976	1.476	.626	.404	.734

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
16.5482	2.864	1.69227	3

Appendix 5

Tests for Normality of Differences of the Two Examiners for the Written Overall Rating

Tests of Normality

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Wr Overall (Ex1-Ex2)	.184	42	.001	.939	42	.026

a.Lilliefors Significance Correction

Appendix 6

Stepwise Regression for Written Overall Score

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	P_Lex Wr	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Ln(Tokens Wr)	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Wr Overall

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df 1	df 2	Sig. F Change
					Change Statistics				
1	.350[a]	.122	.100	.46156	.122	5.572	1	40	.023
2	.474[b]	.224	.185	.43939	.102	5.138	1	39	.029

a. Predictors: (Constant), P_Lex Wr

b. Predictors: (Constant), P_Lex Wr, Ln(Tokens Wr)

c. Dependent Variable: Wr Overall

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.187	1	1.187	5.572	.023[a]
	Residual	8.521	40	.213		
	Total	9.708	41			
2	Regression	2.179	2	1.089	5.643	.007[b]
	Residual	7.529	39	.193		
	Total	9.708	41			

a. Predictors: (Constant), P_Lex Wr

b. Predictors: (Constant), P_Lex Wr, Ln(Tokens Wr)

c. Dependent Variable: Wr Overall

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	4.748	.292		16.244	.000					
	P_Lex Wr	.454	.192	.350	2.360	.023	.350	.350	.350	1.000	1.000
2	(Constant)	-.016	2.120		-.007	.994					
	P_Lex Wr	.541	.187	.417	2.892	.006	.350	.420	.408	.958	1.044
	Ln(Tokens Wr)	.857	.378	.327	2.267	.029	.241	.341	.320	.958	1.044

a. Dependent Variable: Wr Overall

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	Collinearity Statistics VIF	Collinearity Statistics Minimum Tolerance
1	Ln(D Written Data)	.087[a]	.582	.564	.093	.988	1.012	.988
	Ln(Types Wr)	.306[a]	2.159	.037	.327	1.000	1.000	1.000
	Ln(Tokens Wr)	.327[a]	2.267	.029	.341	.958	1.044	.958
	TTR Wr	-.142[a]	-.903	.372	-.143	.893	1.120	.893
	Guiraud Wr	.157[a]	1.034	.307	.163	.956	1.046	.956
	Guiraud Adv Wr	.095[a]	.497	.622	.079	.606	1.650	.606
2	Ln(D Written Data)	.150[b]	1.041	.305	.166	.956	1.046	.927
	Ln(Types Wr)	.155[b]	.747	.460	.120	.464	2.155	.445
	TTR Wr	.117[b]	.604	.549	.097	.540	1.851	.540
	Guiraud Wr	.105[b]	.715	.479	.115	.930	1.076	.903
	Guiraud Adv Wr	.237[b]	1.258	.216	.200	.554	1.807	.554

a. Predictors in the Model: (Constant), P_Lex Wr

b. Predictors in the Model: (Constant), P_Lex Wr, Ln(Tokens Wr)

c. Dependent Variable: Wr Overall

Collinearity Diagnostics[a]

| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | |
				(Constant)	P_Lex Wr	Ln(Tokens Wr)
1	1	1.970	1.000	.02	.02	
	2	.030	8.084	.98	.98	
2	1	2.958	1.000	.00	.01	.00
	2	.041	8.445	.00	.92	.00
	3	.001	74.601	1.00	.07	1.00

a. Dependent Variable: Wr Overall

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.0243	5.9816	5.4167	.23053	42
Residual	-.99261	.69433	.00000	.42854	42
Std. Predicted Value	-1.702	2.451	.000	1.000	42
Std. Residual	-2.259	1.580	.000	.975	42

a. Dependent Variable: Wr Overall

Charts

Histogram
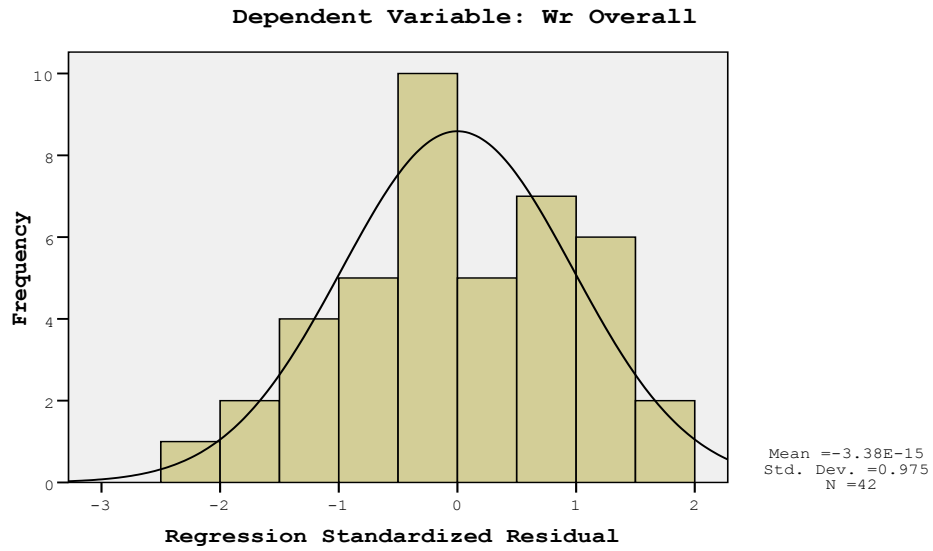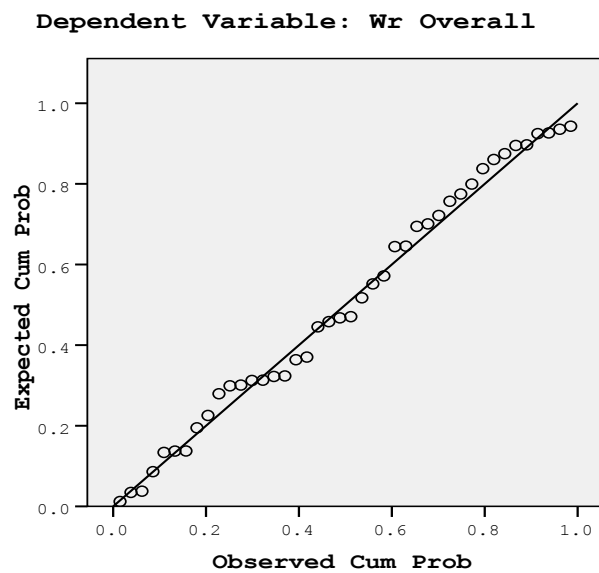
Dependent Variable: Wr Overall

Mean =-3.38E-15
Std. Dev. =0.975
N =42

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Wr Overall

Scatterplot

Dependent Variable: Wr Overall

Partial Regression Plot
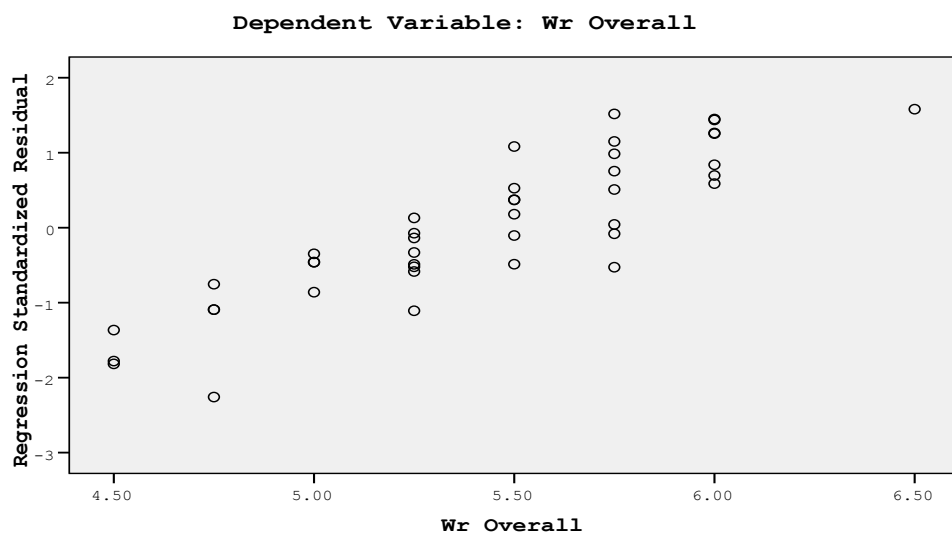
Dependent Variable: Wr Overall
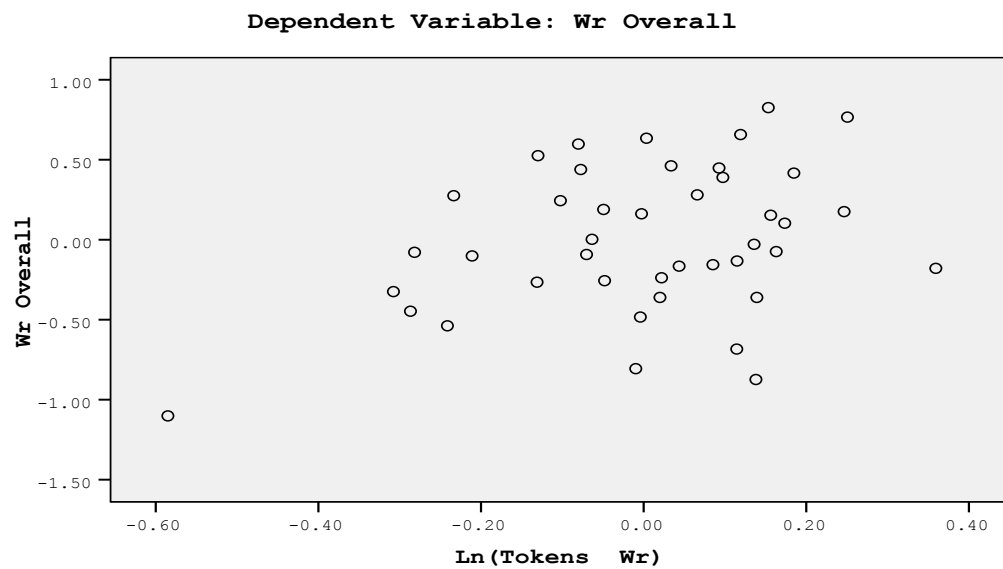
Partial Regression Plot

Dependent Variable: Wr Overall

Appendix 7

Stepwise Regression for Oral Overall Score

Variables Entered/Removed[d]
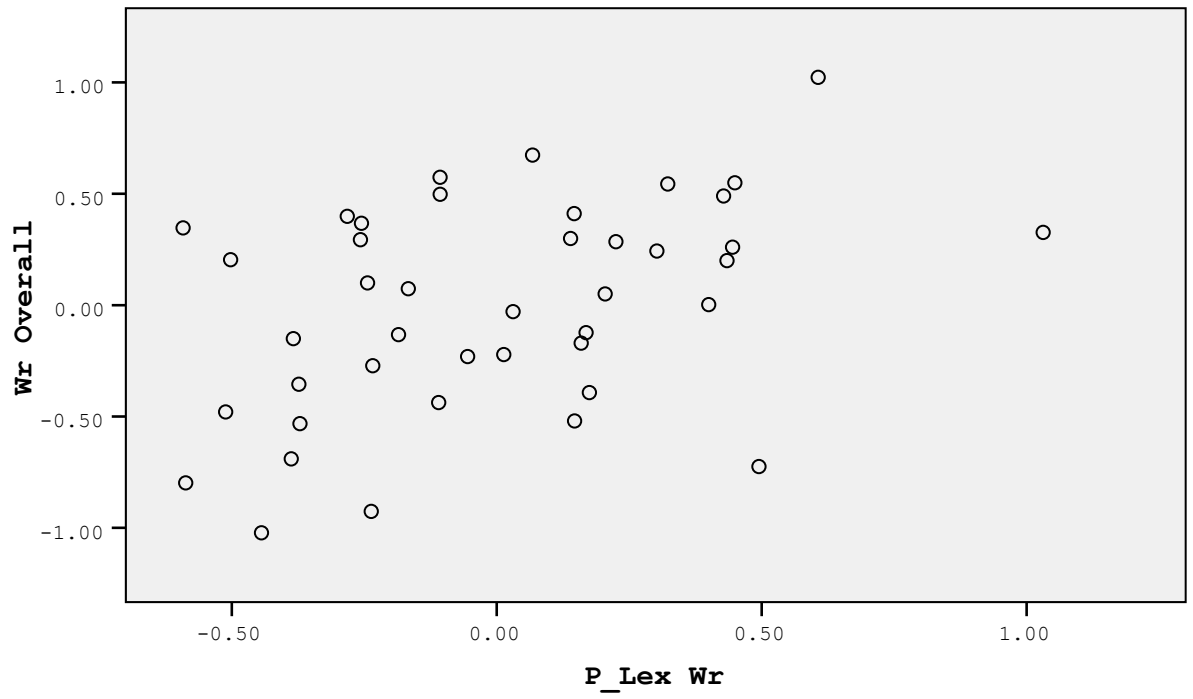
Model	Variables Entered	Variables Removed	Method
1	Guiraud Or	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Or Overall

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df 1	df 2	Sig. F Change
1	.607[a]	.368	.353	.45389	.368	23.325	1	40	.000

a. Predictors: (Constant), Guiraud Or

b. Dependent Variable: Or Overall

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.805	1	4.805	23.325	.000[a]
	Residual	8.241	40	.206		
	Total	13.046	41			

a. Predictors: (Constant), Guiraud Or

b. Dependent Variable: Or Overall

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.845	.763		2.417	.020					
	Guiraud Or	.572	.118	.607	4.830	.000	.607	.607	.607	1.000	1.000

a. Dependent Variable: Or Overall

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		Minimum Tolerance
						Tolerance	VIF	
1	Ln(D Oral data)	-.161[a]	-1.084	.285	-.171	.712	1.405	.712
	Ln(Types Or)	.235[a]	1.005	.321	.159	.289	3.463	.289
	Ln(Tokens Or)	.180[a]	1.039	.305	.164	.528	1.896	.528
	TTR Or	-.130[a]	-.991	.328	-.157	.913	1.095	.913
	Guiraud Adv Or	.077[a]	.609	.546	.097	.998	1.002	.998
	P_Lex Or	.193[a]	1.464	.151	.228	.887	1.128	.887

a. Predictors in the Model: (Constant), Guiraud Or

b. Dependent Variable: Or Overall

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions	
				(Constant)	Guiraud Or
1	1	1.996	1.000	.00	.00
	2	.004	21.751	1.00	1.00

a. Dependent Variable: Or Overall

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.6842	6.3226	5.5161	.34235	42
Residual	-1.01095	.78519	.00000	.44832	42
Std. Predicted Value	-2.430	2.356	.000	1.000	42
Std. Residual	-2.227	1.730	.000	.988	42

a. Dependent Variable: Or Overall

Charts

Histogram
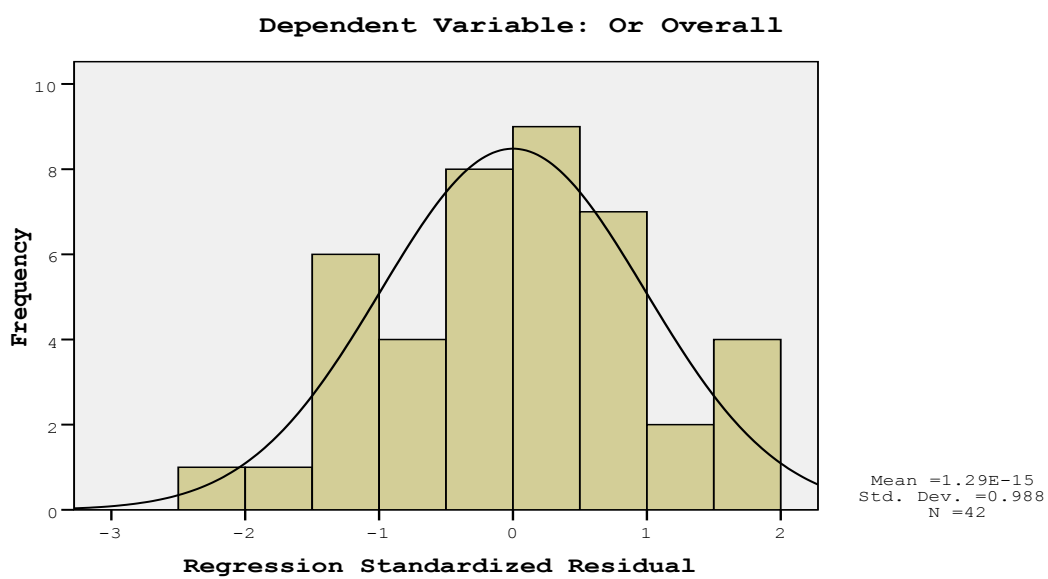
Dependent Variable: Or Overall

Mean =1.29E-15
Std. Dev. =0.988
N =42

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Or Overall
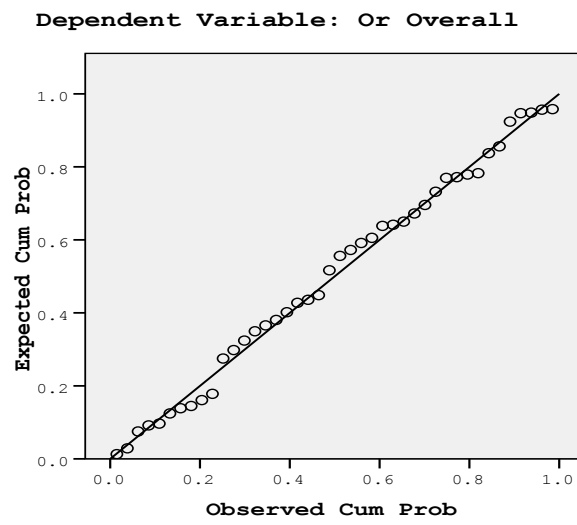
Scatterplot

Dependent Variable: Or Overall
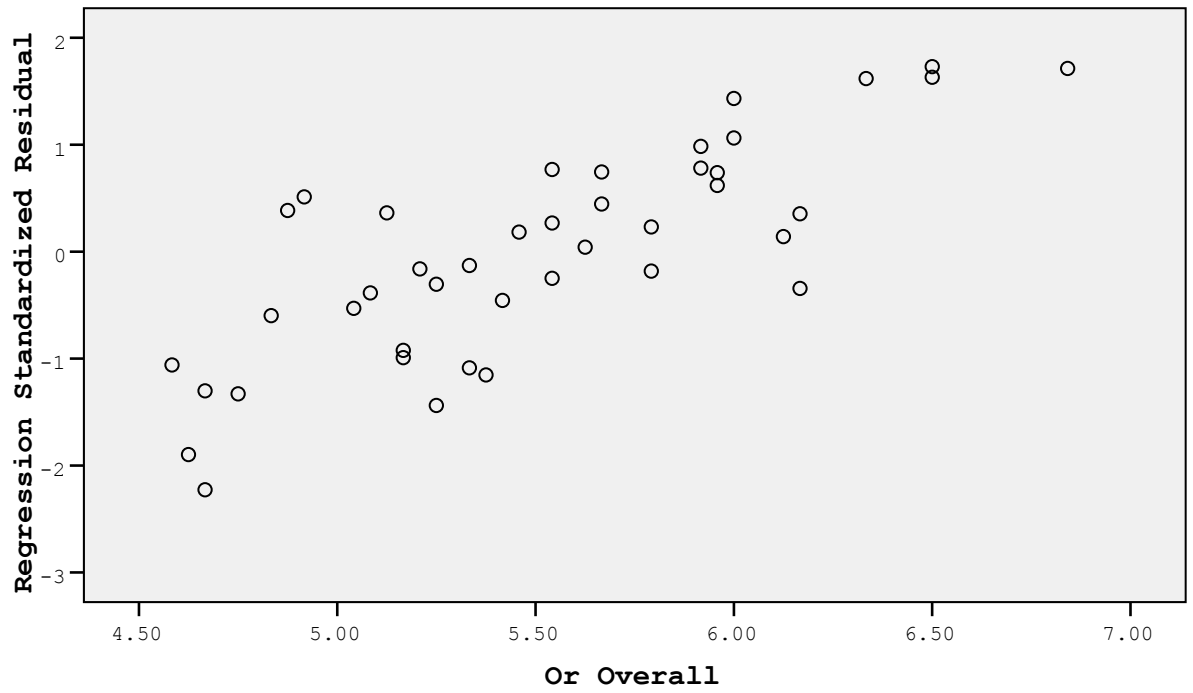
Appendix 8

Martinez and Schmitt (2012) PHRASal Expressions List (PHRASE List)

HAVE TO	WAS TO	UP TO
THERE IS	NOT ONLY	MAXIMUM
THERE ARE	THOSE WHO	A(NY) SINGLE
SUCH AS	DEAL WITH	NO LONGER
GOING TO	LEAD TO	LOOK FOR
OF COURSE	CAUSE	LAST NIGHT
A FEW	SORT OF	AS A RESULT
AT LEAST	THE FOLLOWING	IN ADDITION (TO)
SUCH A(N)	IN ORDER TO	WORK ON
I MEAN	HAVE GOT	THINK ABOUT
A LOT	HAVE GOT TO	FOR INSTANCE
RATHER THAN	SET UP	TOO MUCH
SO THAT	AS TO	YOU SEE
A LITTLE	AS WELL	IN PARTICULAR
A BIT (OF)	BASED ON	A COUPLE OF
AS WELL AS	CARRY OUT	INSTEAD OF
IN FACT	TAKE PLACE	COME BACK
BE LIKELY TO	TEND TO	ON BEHALF OF
GO ON	DUE TO	LOOK LIKE
IS TO	FAIL TO	FIND OUT
A NUMBER OF	EACH OTHER	POINT OUT
AT ALL	IN TERMS OF	APART FROM
AS IF	NO ONE	CALL FOR
USED TO	PICK UP	MANAGE TO

OR TWO	A GREAT DEAL	COME ON
A(NOTHER) FURTHER	ON THE WAY	TAKE ON
COME OUT	AS LONG AS	WORK OUT
BE EXPECTED TO	SO FAR	ALL OVER
SEEK TO	UNTIL NOW	EVERYWHERE
GO THROUGH	OUGHT TO	OTHER THAN
LONG TERM	AT THE MOMENT	TURN OUT
RESULT IN	AS THOUGH	LOOK AFTER
THAT IS	COME TO	AT LAST
EVEN THOUGH	EVOLVE TO	A VARIETY OF
A RANGE OF	ALONG WITH	AT FIRST
THE LATTER	MAY WELL	OR SO
MAKE SURE	COULD WELL	IN FAVOUR
TAKE OVER	GET OUT	IN MIND
CONSIST OF	FOLLOWED BY	GIVE UP
AS SOON AS	IN (THE SENSE) THAT	GET TO
AT THE TIME	THE CASE	ARRIVE AT
WHEN THIS HAPPENED	TAKE UP	FIND ONESELF
ON THE OTHER HAND	ACCOUNT FOR	GET UP
ON ONE'S OWN	SET OUT	CARRY ON
ALL RIGHT	AS FAR AS	GO BACK
SUBJECT TO	CONCERNED WITH	FOCUS ON
AFTER ALL	ABOUT TO	AT ONCE
IN FRONT OF	FIND IT	IT TAKES
TO DO WITH	THINK IT IS	GET ON
GO OUT	SUPPOSED TO	GET OFF
A GOOD DEAL	AND SO ON	AS A WHOLE

IN PRACTICE	SOMETHING LIKE	OVER THERE
BY THE TIME	KNOWN TO	IN SPITE OF
LOTS OF	IN TOUCH (WITH)	THAT'S IT
SAID TO BE	IN THE END	IN PART
IN TIME	IN THE WAY	OH NO
IN TURN	CARE FOR	(WITH) REGARD TO
ONCE AGAIN	IN THE EVENT (OF)	ONE ANOTHER
ALL THE TIME	THEY SAY	AS FOLLOWS
ON THE BASIS (OF)	SO CALLED	THE ABOVE
KIND OF	AT ITS	TO DATE
GET INTO	TAKE INTO ACCOUNT	GO INTO
RELY ON	IN RESPECT OF	TOO MANY
GO FOR	OUT OF	IN THE COURSE OF
AIM TO	AT THE SAME TIME	MORE OR LESS
MAKE UP	NEXT TO	SHORT TERM
APPEAL TO	TURN UP	AIMED AT
END UP	POINT OF VIEW	GO OFF
SHAKE ONE'S HEAD	AT PRESENT	IN CASE
NO MORE THAN	USED TO	OUT THERE
GET BACK	WHETHER OR NOT	LED BY
WHAT ABOUT	IN PLACE	MORE AND MORE
IN OTHER WORDS	NO DOUBT	HAVE A LOOK
AS FOR	FULL TIME	BELIEVE IN
NOT EVEN	SORT OUT	PUT IT
ENTITLED TO	IN A WAY	THESE DAYS
PRIOR TO	OR SOMETHING	IN CHARGE
CHOOSE TO	OR PERHAPS	FEEL LIKE

UP TO	SOMETHING ABOUT	PROVE TO BE
HEARD OF	BY NOW	IN COMMON
TAKE PART IN	THINK SO	NO MATTER
IN SO FAR AS	GO AHEAD	AT THIS POINT
PART TIME	BRING ABOUT	IN ITSELF
LOOK FORWARD TO	HAD BETTER	THE FORMER
AS SUCH	IN ACCORDANCE WITH	IF ONLY
BOUND TO	CALL ON	YET TO
TURN ON	AT TIMES	UP TO (DECISIONS)
SET TO	ALL THE WAY	OR WHATEVER
MOVE ON	IN EFFECT	HAND OVER
IN CONTRAST (TO)	AFFORD TO	IN THE LIGHT OF
THIS STAGE	SIGHT OF	IN THE SAME WAY
ALL BUT	IN ADVANCE	THAT MUCH
ABOVE ALL	ON THE PART OF	THE EXTENT TO WHICH
RID OF	BRING UP	FOR SOME TIME
IN ANY CASE	TAKE OFF	IN RETURN (FOR)
THANKS TO	SO AS TO	TO DEATH
GO AWAY	TAKE ADVANTAGE	ON THE GROUNDS
ONCE MORE	SHORT OF	OH DEAR
OH WELL	OVER THE YEARS	IN FULL
FOLLOW UP	SWITCH ON	ON BOARD
WOULD SAY	BY NO MEANS	TO SOME EXTENT
FOUND TO	COULD HARDLY	SOME KIND OF
MEANT TO	COME UP WITH	KEEP UP
HANG ON	IN QUESTION	NO IDEA
TURN INTO	IN THE FIRST PLACE	GREATER THAN

HAPPEN TO (BE)	GET ON WITH	ADD TO
HELD THAT	NO GOOD	AS YET
FACED WITH	YET ANOTHER	AT RISK
DO(ING) SO	KEY TO	A MERE
SET OFF	I'M AFRAID	SHOWN TO
PUT FORWARD	THAT WHICH	ON THE ONE HAND
FROM TIME TO TIME	IF SO	BY WAY OF
THE MEANS	RIGHT NOW	ON THE ROAD
EVER SINCE	IN VIEW OF	OLD FASHIONED
JUST ABOUT	IN DETAIL	FOR SALE
GIVE RISE TO	REFLECTED IN	OR ANYTHING
LARGE SCALE	NO SUCH	MOST LIKELY
MAKE SENSE	NOTHING BUT	PROVIDE FOR
BY MEANS OF	IN THE FACE OF	EVEN SO
IN SHORT	SUCH THAT	COME ACROSS
A BIT OF A	NEXT DOOR	NO FURTHER
BREAK UP	TO THE POINT	FIRST OF ALL
ALL TOO	MAKE ITS WAY	MIGHT AS WELL
PUT UP	MAKE ONE'S WAY	LIMITED TO
GOOD AT	IN HAND	TO ME
A LONG WAY	PARTY TO	IN MY OPINION
AMOUNT TO	BY THEN	MIND YOU
FOR LONG	GET TO	AT A TIME
(BE) RUN BY	BY THE WAY	HALF PAST
SOME MORE	BY CONTRAST	WITH RESPECT TO
IN THE ABSENCE OF	RUN OUT (OF)	CONSISTENT WITH
ALL SORTS OF	IN PRINCIPLE	WAY OUT

THIRD PARTY	FOR ALL	IN THEORY
CONTRARY TO	A QUESTION OF	THOUGHT OF (AS)
WORTH OF	FOR LIFE	FOR GOOD
A GOOD	GET AWAY	OPPOSED TO
AT LEAST	IN THE MEANTIME	COMMON SENSE
ACT ON	SOMETHING OF A	BOTHER TO
EXCEPT THAT	THE ODD	AS GOOD AS
DAY TO DAY	LITTLE MORE THAN	BACK UP
AS USUAL	WOULD YOU LIKE	TAKE CARE OF
LONG BEFORE	IN NEED	THE SIGHT OF
LONG AGO	TAKE FOR GRANTED	GO ROUND
IN CONJUNCTION WITH	IN THIS RESPECT	THE WHOLE THING
UP TO DATE	PROVIDED THAT	AT ONE TIME
LET ALONE	ALLOW FOR	HEAD TO
QUITE A LOT	CALCULATE IN	IN A SENSE
IF YOU LIKE	CATCH UP	ON AVERAGE
TO THE EXTENT	A GO	WAY ROUND
SO FAR AS	FOR THE MOMENT	CAN TELL
GIVEN THAT	AT THE EXPENSE OF	FREE FROM
IN LINE WITH	PUT TOGETHER	AND ALL THAT
ON THE WHOLE	THINGS LIKE THAT	AS IT WERE
CARE TO	OF LITTLE	TOUCH OF
TAKE ACCOUNT OF	SHUT UP	BETTER OFF
SOMETHING LIKE THAT	AS OF	STAND FOR
MAKE USE OF	OVER TIME	TO BLAME
WHEN IT COMES TO	WOULD APPEAR	THE BULK OF
FILL IN	THE OTHER DAY	A HANDFUL OF

(AT) THE OUTSET	FOND OF
BY VIRTUE OF	WITH A VIEW TO
TURN DOWN	TURN BACK
GET ON	GET AWAY WITH
UNDER WAY	NO WONDER
IN THE INTEREST OF	WELL BEING
ON THE MARKET	HOW ABOUT
BY FAR	TO GO
A DEGREE OF	STRAIGHT AWAY
NEVER MIND	OWING TO
UP AND DOWN	HOLD UP
IN ONE'S OWN RIGHT	LOOK TO
A CASE OF	LAY OUT
MORE SO	THE LOT
COME UP TO	KEEP ON
IN WHICH CASE	MAKE UP ONE'S MIND
NO SIGN OF	AT WORK
JUST AS	COME ABOUT
FOR THE SAKE OF	
IN A POSITION TO	
TO COME	
BACKED BY	
AT BEST	
WEALTH OF	
THAT SORT OF THING	
MAKE OUT	
COME TO TERMS WITH	

Appendix 9

General Procedures for Text Preparation

Notes

The purpose of these procedures was to ensure, as far as possible, that the texts were uniform in respect of certain features, to preserve the integrity of any statistical analyses.

It is accepted that such procedures are always open to argument and can always be approached differently.

Explanatory notes for some of the procedures have been included.

Preparation of Texts for VOCD

Proper nouns – deleted.

Spelling – corrected.

Incomprehensible words neither removed nor corrected.

The problem here was intentionality and ambiguity. Such a 'word' would not count as a token but the writer obviously intended a particular meaning, or a 'type'. It was thus considered appropriate to recognise the intentionality and preserve its status as a token and as a Not-on-List word.

Misused/incorrect words – deleted.

Articles joined to following words – separated.

Text reference numbers and identification marks – deleted.

Numbers written as words in American expression (e.g. *'two hundred twenty'*).

Some numbers were written as numbers, some as words and some, a mixture of both. The omission of 'and' is in keeping with the teaching of the American expression that, at the time, was standard.

'12 am' and *'12 pm'* written as *'twelve midnight'* and *'twelve midday'* respectively

'a.m.' and *'p.m.'* referred to as 'morning', 'afternoon' or 'evening' as appropriate (e.g. 'two morning' or 'five afternoon').

Other times (e.g. *at 2.30*) changed to word form (*at two thirty*) with no reference to a.m. or p.m. if it is clear from the context, otherwise the above applies

Proper nouns used as adjectives – deleted.

Abbreviations – deleted.

Duplicated words – one deleted (e.g. *the the …*).

Misplaced apostrophe (e.g. genitive, contractions) – corrected.

'to' – corrected to *'too'* if context requires.

Words in isolation – deleted.

Non-English words (e.g. *souk/souq*) or foreign words not usually used in English – deleted.

Referencing accepted as: *'according to listening/reading'* or only *'listening'* or *'reading'*. *'L'* and *'R'* changed to word form. Any expansion (e.g. *the* listening) *'the'* deleted.

Referencing is a requirement in writing but options as to how to reference are permitted. This leads to anything from a single letter ('L' if it is from the listening text or 'R' if from the reading) to one token (simply 'reading' – to indicate the reference was from the reading text) to more tokens ('from the reading').

Quotations properly included and relevant to text – accepted.

Quotations improperly included or irrelevant to text – deleted.

In the lower levels more than the upper, students, either intentionally or, perhaps through misunderstanding, use a *quotation out of context, thereby creating ambiguity.*

Per cent – number written in full with words 'per cent'.

'themself', *'hisself'*, *'theirself'* etc. – regarded as spelling errors and corrected.

'responsibilitys', *'babys'*, *'womens'* etc. – regarded as spelling errors and corrected.

Dates – month deleted, date written as number.

Text Identification

Student's University ID + essay reference number.

Raters' Essay Copies

All plain text, single spaced, unedited, no identifying marks or dates apart from reference number. One copy per rater.

Rating

Prior meeting to confirm all requirements, independently rated, no conferring on any result, texts randomly allocated, no time limit enforced. IELTS criteria used. Raters may write on their copies if they wish.

Raters to regard each essay as an entry for an IELTS examination.

Only vocabulary and holistic rating required.

IELTS criteria and marking scale used.

Appendix 10

P_Lex Analysis- Words that did not belong in Level 1 list

Student 1

Active, adopts, adopted, adoption, aid, appropriate, benefit, betray, citizen, code, communicate, companies, computer, conclusion, conflict, culture, cultural, desire, disagree, disaster, discriminate, essay, experts, extinct, finally, firstly, fulfill, future, globalization, healthcare, heart, high-status, household, humanitarian, industry, involved, kid, loyalty, marine, normal, nowadays, organize, orphan, ozone, personality, pollute, preparatory, primary, replace, responsible, robots, secondary, society, solutions, solve, stage, starvation, supervisor, tankers, third, thirdly, tradition, training, type, unfortunately, upset

Student 2

Absent, activities, advantages, balance, basic, business, cheaper , communicate, company, conclusion, conflicts, consider, consult, day by day, deal, defusing, directly, disadvantages, discrimination, educated, economic, eighty, empathy, encourage, endangered, essay, extinct, exploration, evaluate, factors, faster, fifty, final, firstly, fist, future, gentle, globalization, goal, helicopter, hiring, household, identify, located, loss, million, misunderstand, model, native, negative, offer, outcome, overseas, oxygen, particularly, past, professional, promotion, qualifications, reduce, resolve, responsible, rates, reduce, rethink, returned, robot, salary, secondly, socially, solutions, solve, statements, stroking, sum, summary, technology, therefore, third, thousands, tongue, training, tribes, unemployment, view

Student 3

According, adoption, adventure, agency, afford, aid, apart, appropriate, areas, asthma, aware, basic, cancers, capital, climate, conclusion, couples, crises, deal, desert, dinner, directly, disaster, disagree, distributed, earthquake, encourages, equipment, essay, explored, finally, fisherman, fixtures, flexibility, forty, funding, funny, generation, headache, healthcare, home, homework, household, huge, humanitarian, hundred, industry, insure, international, luggage, major, marine, monetary, moreover, natures, nineteen, occur, opportunities, orphans, oxygen, parentless, past, phone, physically, pollute, population, projects, promote, provided, relax, responsibility, respite, research, return, salaries, sale, same, search, shelter, similarity, solution, statement, spread, tankers, third, threatened, text, thousand, topic, toys, tradition, transport, trips, type, variety, victims, whatever

Student 4

Able, according, adventures, adopted, affect, aid, airport, appropriate, arrived, balance, benefit, bound, business, camping, career, chair, challenges, channels, computer, company, conclusion, couples, culture, customer, dinner, disasters, dreamland, endangered, environment, essay, extinct, fantastic, fifteen, finally, firstly, flexible, full-time, grades, graduate, hiring, homeless, humanitarian, identity, immigrate, improve, informal, interview, lashes, lunch, mole, museum, nap, native, negatively, outgoing, outsourcing, overseas, part-time, past, perfumes, qualify, refugees, remain, responsibilities, secondly, shower, social, solutions, solve, standard, stakes, strategy, suggestion, summary, supermarket, support, survive, tents, text, third, thirty, thousand, traditional, trip, type, unbelievable, untraditional, wonderful

Student 5

According, adoption, adults, affecting, aid, alive, appropriate, areas, arrive, assessments, background, balance, boss, camels, camp, cases, celebrating, chess, consumer, conclusion, cooperate, culture, desert, diagnose, endanger, essay, falcons, final, firstly, free-trade, future, globalization, grandmother, guide, high-stakes, humanitarian, improve, indication, intend, interrelated, issue, kid, lunch, majority, mount, nap, non-traditional, offers, omnipresent , orphanages, outcome, outgoing, overcome, overwhelming, past, professional, pupils, regional, responsible, score, seventeenth, shower, solutions, standardized, stuck, sum, tent, third, thirty, thousand, twenty, type

Student 7

Activities, advantage, adopt, airport, ancient, attract, aware, benefit, billion, boring, communication, companies, computer, culture, deal, delays, dollars, economies, emotional, encourage, endangered, extinct, factors, fare, fifty, finally, focus, gifts, globalization, hero, huge, hundred, impact, inter-relationships, million, minimize, newspaper, nowadays, paragraph, percent, physical, population, programs, replace, researchers, robots, seventy, ships, solutions, strict, survive, team, tongue, trip,

Students 8

Abilities, achievement, administer, adopting, arrived, attention, barriers, career, communication, conflict, contacts, cultures, dinner, disaster, discrimination, donkey, encouragement, enter, environment, explorer, finally, fine, flexibility, framework, fun, future, garbage, goal, hen, however, hunting, ill, independent, informal, intelligent, juggle, lack, lecture, legal, levels, major, moreover, native, negative, oxygen, phoned, planets, popular, professional, promotion, protect, provided, relax, responsibilities, rates, role, salary, sector, sex, ship, sheep, sites, stakes, support, suppose, therefore, third, wonderful,

Student 10

Adopt, affects, article, attracted, belong, bug, calculate, cartoon, civil, cleansing, comment, communicate, connect, consumer, crossed, cultures, delicious, disappear, disaster, diversity, dominate, earthquake, economic, endangered, ethnic, fans, fast, favorites, first, fries, generation, globalization, international, involved, lecturer, literacy, literature, maintain, media, multinational, offer, opponents, past, phenomena, poverty, process, programmes, reduce, refugees, resist, selection, silly, social, sorting, sources, spice, spread, strive, survive, theatre, toy, traditions, transformed, variety, video, wonderful,

Student 11

According, adventure, aids, ancient, balance, basic, beaches, bored, camels, citizen, club, colonization, communicate, conclusion, consider, cucumber, cultures, disappear, disasters, discrimination, dominated, endangered, essay, exploring, extinct, finally, firstly, goats, graduate, heaviness, huge, improve, information, injured, juggle, kids, loneliness, marine, moments, negatively, nervous, nineteen, ninety, ocean, official, pesticide, protect, qualify, remains, responsibility, resources, return, samples, sheep, shelter, ship, shrub, split-shift, spread, survive, third, thousands, transports, trip,

Student 12

According. adoptions, advantages, aid, betrayed, bleeding, checking, communicate, couples, conclusion, conflicts, consumer, decrease, disagree, disasters, divorce, dominate, encourage, endangered, earthquake, essay, extinct, finally, firstly, future, globalization, graduate, hundred, investment, maid, marine, mention, negative, nowadays, organize, percentage, provide, recognize, responsibility, salary, scared, shelters, shift, shy, solutions, spread, survivor, tankers, thirdly, thousand, training, tsunami, unique

Student 13

Activities, airport, Arab, arrived, average, barriers, beach, beat, boss, breakfast, capital, camping, ceiling, charge, chess, dinner, disagreement, equipment, emotional, fifty, first, flexible, forty, hard-working, hometown, housemaid, humanitarian, ill, inflexible, involved, juggling, lecture, legal, loyal, lunch, malls, mention, midnight, misuse, museum, nap, newspaper, percent, population, quote, returned, sector, shower, social, supportive, taboo, tents, thirty, tour, towers, trustworthy, type, views, waterfall,

Student 14

Adopt, appeal, appropriate, backpack, beside, camels, cans, centered, crises, computer, chat, crabs, communication, companies, cool, culture, dawn, dealing, delicious, diversity, drought , eel, endangered, environment, famine, favorite, final, flood, foolish, global, graduated, homeland, handsome, housewife, import, injures, letting, loss, messy, million, ocean otherwise, overseas, past, pension, pool, prompt, rebuild, region, relax, remain, removing, replace, research, resource, retires, roadways, robot, score, shellfish, solve, specific, spill, standardized, turtles, tanker, technology, temporary, twenty, visual, volleyball, within, wonderful, zoo

Student 15

Ability, according, adult, advanced, affective, alive, ambition, basic, challenge, chemistry, communicate, cope, disaster ,dynamite, economy, emirates, encouraged, endanger, environment, feed, fun, funny function, healthcare, hobby, industry, immigrate, intelligent, instruction, lack, located, marine, man-made, media, million, native, pupil, rebuild, recovery, replace, review, robots, sexes, shelter, spills, stereotypes, telecommunication, view

Student 16

Active, airport, apart, aqua, amazing, assessments, benefits, blast, characters, century, colleges, companies, conclude, confidence, courses, delicious, disadvantage, discrimination, drift, earthquakes, eighty, enter, explorer, floods, focus, glad, grammar, hit, higher-level, immigrants, income, levels, lifestyle, low-level, moments, moreover, nets, ninety, offer, organizes, paradise, particularly, perfectly, poverty, prevent, prize, professional, project, promotion, providing, rebuild, recent, responsible, restrictions, role, samples, sand, semester, ships, solutions, specifically, stakes, tankers, theory, therefore, traditional, upset, vital

Student 17

Able, abilities, according, achievement, actual, adapt, adoption, adventure, affected, aid, anxious, appeal, apply, appropriate, areas, association, asthma, available, avoid, balance, basic, barking, biodiversity, basic, centers, climate, college, companies, conclusion, contain, crisis, define, disaster, earthquake, economic, eighteen, enter, enveloped, explorers, finally, firstly, focus, football, fund, gallery, globalization, habitat, headphone, however, humanitarian, impact, improve, indicate, industrial, infer, influences, information, international, kids, kilometres, lack, location, major, medical, migrate, motivate, native, negative, nineteen, ninety, nowadays, nuts, overall, percent, phenomenon, pollution, poverty, professor, protect, provides, reconstruct, refugees, raise, recycling, role, shelter, sixteen, solved, solutions, split-shift, spread, standardized, suggestion, summary, survey, survive, tennis, terrific, thousand, traditions, transportation, tourism, trip, type, wondering, vital

Student 18

Achievement, active, adoption, adults, advantage, affected, aid, appropriate, areas, authorities, balance, blast, cases, college, conflict, coral, companies, computer, conclusion, create, deal, decrease, disaster, drift, dynamite, endangered, enforce, essay, export, extinct, finally, globalization, habitat, hit, huge, hundred, humanitarian, immigrate, improvement, individuals, insurance, international, interview, item, levels, marine, negative, nets, newspaper, nowadays, outsourcing, overseas, past, pollute, population, protect, publish, raise, rates, reefs, remain, report, resolve, scores, seventy, spills, standardized, support, tankers, third, thousands, tongue, type, video, view

Student 19

Ability, absent, according, activities, administrator, advanced, advantage, affect, aid, ancient, balance, basic, benefits, biology, billions, capitals, century, charge, chemistry, colony, computer, communicate, companies, concerns, conclusion, constructions, decrease, discrimination, distribute, economic, encourage, endangered, experts, exploited, extinct, fifty, finally, fourth, globalization, heart, healthcare, higher-status, hiring, hundred, immigrate, import, improve, individuals, labour, lack, lecture, levels, located, major, malaria, millions, moreover, nowadays, official, orphanage, outsourcing, palaces, past, port, poverty, protect, receive, reduce, remain, remove, researchers, responsibilities, shelter, social, soldering, solve, solution, starvation, stereotypes, suggested, supervisors, supportively, survey, telecommunication, therefore, third, thousand, transportations, trip, twenty, type, view, welcome, wheat,

Student 20

Ability, advantages, addiction, aid, appeal, attract, balance, career, chemistry, collect, communicate, conclusion, conflict, consider, cool, co-operative, crises, cultural, deal, decrease, disaster, disadvantages, economic, emotion, enhance, environment, essay, exercise, extinct, fast, finally, football, formal, gains, handball, hobbies, homemade, homemaker, humanitarian, hundred, improve, individuals, informal, interpersonal, lack, landfill, loans, loyalty, micro, millions, negative, ninety, past, percentage, poverty, protect, rebuilt, relax, responsible, restrictions, returned, sector, shame, shelter, social, solutions, tankers, tennis, third, tons, traditional, unable, volleyball, wonderful

Student 21

Able, according, adoption, aid, appropriate, arrived, astonished, avoid, balance, basic, betrayed, biodiversity, bosses, ceiling, challenges, coastline, collect, companies, communicate, conclusion, decrease, delicious, dominated, dynamite, eighteen, emotional, encouraging, enter, environmental, essay, essential, fifty, fifteen, finally, first, flexible, future, grade, guide, homework, homes, huge, hundred, humanitarian, interviewed, issues, lack, levels, longed, mail, majority, male-dominated, million, negative, ninety, nineteen, outside, percentage, prevent, promotion, protect, pupils, requirement, relax, responsibility, returned, roles, salary, score, sector, sexes, shelters, ship, solution, source, species, spilling, standardized, stress, suggested, supermarket, support, tankers, third, thirty, toys, transport, type

Student 22

According, adoption, affect, airport, alive, amazed, bat, beach, belonging, benefits, breakfast, cans, chairs, chess, company, computer, conclusion, creatures, culture, deserts, dinner, disasters, endanger, enter, environment, essay, extinct, feed, finally, firstly, focus, forty, fun, globalization, graduation, hard-working, hit, homeless, housemaid, identity, information, injures, jellyfish, lunch, major, man-made, marine, monkey, museum, negative, nowadays, ocean, opportunity, option, orphans, outline, pacific, past, perform, pollute, pool, populations, raise, rarely, relax, responsibly, robots, schedule, scorpion, search, solution, stakes, sum, tanker, tent, third, thirty, toys, traditional, turtles, twenty

Student 23

Ability, according, adoption, advantages, aid, available, basic, beard, billion, bills, boost, cartoon, channel, company, conclusion, consumers, convince, couples, deal, deal, disaster, doll, dynamite, earthquake, eighty, encourage, enforce, environment, fifty, finally, firstly, flexible, focus, future, globalization, huge, humanitarian, lack, marine, media, movie, negative, ninety, nowadays, organize, population, protect, pupils, raise, region, replacement, responsible, returned, robot, roles, shelter, smart, solve, solution, spills, stereotype, tankers, theme, thirdly, training, type, view

Student 24

According, achieve, adoption, affect, aid, appropriate, balance, basic, blast, career, ceiling, colonization, combine, compulsory, concentration, conclusion, coral, culture, curriculum, disadvantage, discrimination, dynamite, endangered, enroll, enter, environmental, essay, extinct, factor, final, flexible, fortunately, future, humanitarian, international, lecture, marina, million, movie, multimedia, negative, non-traditional, nowadays, object, optional, organize, permission, polluters, population, poverty, reef, remove, responsibility, role, salary, search, shift, solve, solution, specific, spread, spill, stakes, standardized, summary, supply, tankers, therefore, third, tourism, transport, trip, type, view, zoo

Student 25

Abilities, according, advantages, adoption, aid, appropriate, areas, articles, balance, barriers, boss, ceiling, coastal, college, communicate, companies, computer, conclusion, considered, contribute, coral, crabs, culture, demonstrate, disaster, discrimination, eels, encourage, enforce, enormous, environment, essay, extinct, favorite, finally, flexible, gallons, globalization,

graduate, high-stakes, home, humanitarian, hundred, hit, importance, information, item, juggling, laundry, lecture, levels, luxury, million, observe, occupy, past, pollute, prevents, professions, programs, protect, qualifications, relaxed, raise, recent, recognize, reefs, researchers, salaries, score, senior, services, solve, solutions, source, standardized, stress, strict, survive, tanker, therefore, third, thousand, training, transport, type, vital

Student 26

Able, accountable, adoption, affect, aggressive, agriculture, aid, airport, anonymous, apple, appropriate, aware, balance, basic, blast, circumstances, communication, conclusion, conflict, conversation, co-operate, couples, create, crises, deal, decrease, disappear, disasters, dynamite, earthquake, emotional, endangered, endure, essay, excluded, extinction, factors, finally, firstly, floods, future, highlight, homework, horrible, however, humanitarian, hundred, ignore, improve, income, individuals, insecure, likewise, lockers, loyalty, lunch, major, marine, mention, moment, moreover, negatively, normally, nowadays, opportunity, orphanage, past, peer, physically, pollution, programs, protect, receive, remains, resolve, role, qualifications, solution, solve, surrounding, survive, tanker, tasks, therefore, third, thousand, thus, topic, traditional, types, uniform, verbally, victims, wonderful

Student 27

Able, adopting, advantages, affect, agencies, aid, appropriate, aware, balance, barriers, basic, boss, career, chores, classify, companies, computers, conclusion, considered, coral, couples, culture, disadvantages, disaster, domestics, dynamite, economy, emotional, endangered, energy, essay, ethical, ethnic, exit, experts, extinct, fascinating, faxes, finally, financial, future, globalization, goal, grandparents, healthcare, hobbies, however, humanitarian, improve, individuals, industries, international, marine, million, native, nowadays, occur, opportunity, past, prevent, protect, protein, raise, reefs, reflects, removing, salary, sector, solve, solution, standard, strategy, supportive, tanker, technology, therefore, third, thousands, type, variety

Student 28

Able, according, adapt, adequate, adoption, adult, advancement, aid, alive, appropriate, areas, bases, basis, balance, benefit, cases, ceiling, chores, committed, company, conclusion, consequences, corrupt, disconnect, cultures, continuous, divorced, dominate, disadvantages, economical, emotional, entering, environment, essay, eventually, expression, exemplify, extinct, final, fulfill, fun, future, gap, garbage, globalization, harm, homeland, humanitarian, illustrate, improve, income, infringement, interact, items, jealous, jointly, joy, lacking, links, loyalty, merge, millions, misconception, moreover, mutual, negative, nowadays, opportunities, orphanages, oxygen, past, professional, profit, promoted, proposes, protect, provider, quote, raise, recently, recommendation, sale, secretary, sector, seek, sensitive, ship, similarly, states, solutions, solving, sorts, source, statement, stranded, suggests, supporter, tend, theory, thousands, tissue, tradition, trip, type, wealthier, workforce

Student 29

According, adopting, affected, aid, borders, cases, chair, chemical, communicate, company, conclusion, conflict, current, deal, defusing, disadvantages, discriminate, drug, earthquake, encourage, endangered, environment, essay, extinct, finally, firstly, flood, fossil, funny, future, glad, global, globalization, hit, huge, humanitarian, hundred, information, let, moreover,

nowadays, picnic, protect, pupils, refresh, responsible, salaries, seals, silly, smarter, solutions, solve, tankers, third, thousand, transport, type, view

Student 30

Able, according, adoption, affect, afford, agriculture, aid, ancient, article, balance, basic, beneficial, boring, borrow, chair, civil, communication, company, confident, conflict, confront, conclude, conclusion, crises, culture, cutest, detrimental, disadvantages, disaster, divorce, earthquake, economic, endangered, enrolling, essay, expand, extinct, finally, firstly, flood, funny, future, globalization, guide, heart, huge, ignore, illness, issues, loan, moreover, nowadays, nutrition, occupation, occurs, past, promotion, pupils, recently, reconsider, resolve, responsibility, return, role, sacrifice, scared, search, security, shelter, solution, solve, support, survive, therefore, thus, variety, view, wealth

Student 31

According, active, adoption, aid, appropriate, areas, background, barrier, beach, benefit, biology, blast, colleges, communicate, companies, computer, conclusion, conflict, conversation, crises, culture, deal, decreased, diagnose, discrimination, drought, dynamite, earthquakes, endangered, enter, essay, except, extinct, feather, final, firstly, floods, focus, future, generation, geography, globalization, grade, grammar, home, humanitarian, hundred, improve, lecture, majority, marine, mention, million, moreover, newspaper, ninety, oral, pollution, programs, projector, pupils, quit, replace, responsible, robots, salaries, sacrifice, score, search, section, specific, spills, solve, solutions staff, stakes, starvation, stereotype, survived, swallow, tanker, third, thousand, training, type, view

Student 32

According, advance, adventure, aid, airport, assistance, average, avoid, balance, basic, barriers, beaches, billion, capital, company, computer, conclusion, conflict, create, crises, deal, desert, disaster, discrimination, dominated, emotional, encourage, environment, erosion, essay, explorer, explosive, finally, financial, firstly, flexible, formal, function, future, healthcare, home, household, hundred, improve, individuals, industries, intelligent, lack, layer, let, micro-financial, moreover, nowadays, opportunities, oxygen, ozone, participate, percentage, promoted, protect, provided, pupils, reduce, requirement, resolve, responsibilities, sector, senior, shelter, ship, small, solution, solve, stream, styles, supplying, support, survive, therefore, third, thus, trip, type, vary, victims, warning

Student 33

Able, abuse, according, aids, airlines, alive, ambitions, ban, benefit, boring, career, companies, communication, confide, connection, construction, create, crises, culture, disappearing, discrimination, dynamite, emotional, encourage, endangered, enforce, essay, finally, formal, fund, globalization, healthcare, homemade, homework, however, individuals, informal, international, labor, lack, marine, native, ninety, normal, nowadays, oral, palaces, past, paragraph, percent, pollute, providing, programme, protect, raised, responsible, salaries, search, sector, sexes, shy, solving, solutions, smaller, smart, specific, spilled, spreads, starvation, strict, summary, support, third, training, type

<u>Student 34</u>

According, adult, aid, average, balance, basic, benefit, billion, company, conclusion, create, cultures, currently, deal, deposit, dissimilarity, distributed, dollars, economic, endangered, essay, enforce, finally, globalization, healthcare, homes, individuals, majority, millions, nervous, nineteen, ninety, nowadays, outsourcing, overseas, percent, pollution, population, poverty, projects, providing, pupils, receive, reduces, salaries, ships, solutions, solve, spread, standard, starvation, strategy, summary, technological, therefore, third, threat, view, wealth, wheat withdrawal, worth

<u>Student 35</u>

According, aid, assist, avoid, balance, beach, camel, capital, clinic, communication, company, complex, conclusion, conflict, connect, counties, culture, deal, disaster, encourage, endangered, environment, essay, essential, extinct, extremely, finally, flexible, future, healthcare, heart, home, improve, interpersonal, loan, major, marine, micro, nowadays, opportunity, pollution, protect, resolve, response, responsibility, return, solution, solve, tankers, third, thousand, tourist, transport, unique, view

<u>Student 36</u>

Abilities, abroad, according, agriculture, aid, available, average, avoids, balancing, basic, breeding, certificate, ceiling, center, companies, create, culture, deal, destruction, discrimination, earthquake, encourage, endangered, environment, epidemic, essay, extinct, filter, finally, floods, grade, habitat, homeland, homes, hunt, illness, improve, individuals, interface, interference, invite, lack, lecture, liberate, lonely, luxury, marine, media, nowadays, opportunities, palace, past, percent, population, poverty, promotion, protect, receive, reduce, responsible, salaries, select, senior, solution, solve, spills, spread, starvation, stereotype, survive, third, thousand, tongue, tradition, training, vitamin

<u>Student 37</u>

Accident, according, achieve, aid, airlines, arises, arrived, attraction, aware, beach, benefit, career, certificate, comments, computer, conclusion, conflict, construction, college, culture, deal, disaster, emails, encourage, essay, experts, expressing, expose, famine, finally, finance, firstly, fund, funny, furthermore, gap, harmful, homework, ignoring, industry, interpersonal, impact, income, instead, insult, intelligent, legal, loyalty, lunch, major, mutual, negative, nowadays, palaces, past,   percentage, promotion, protect, provides, receive, reputation, responsibly, resources, role, salary, shelter, ships, shy, solve, statement, stress, struggle, summary, support, tankers, third, tourism, view, vital, wonderful

<u>Student 38</u>

According, achieving, aid, alive, apartments, arrived, basic, benefits, boring, career, centers, citizens, companies, conclusion, conflict, decreasing, destruction, discrimination, economy, emotional, encourage, environmental, essay, explorer, finally, flexible, formal, glad, grief, home, illness, illogical, informing, improve, instead, intelligent, interpose, joy, lack, lonely, misery, monsoon, nineteen, organization, outlined, overcome, percent, planet, pollute, prevent, privilege, protect, providing, qualification, recently, reduce, relaxation, responsibilities, role,

salary, seeking, sexes, ship, solutions, solve, sources, stereotypes, survive, technology, therefore, third, traffics, training, transportation, trash, view, weapons

Student 39

Abilities, active, addicted, adopted, afford, aid, appropriate, available, balance, ban, basic, boss, breakfast, canteen, challenges, chatting, companies, computer, connected, conclude, consequences, consider, consist, consumer, counties, couples, create, culture, desert, disadvantages, earthquake, economy, encourage, enter, essay, facilities, firstly, finally, formal, graduate, grammar, globalization, homemade, huge, humanitarian, ignore, import, influences, informal, internet, jail, lack, lecture, linked, motorbikes, nowadays, negative, opportunity, optional, outline, past, percentage, pregnant, prevent, principal, programs, project, reminder, replaced, responsible, role, salary, sales, schedule, sector, self-esteem, shower, site, social, solve, solution, specific, stake, standardized, stereotypes, summary, support, therapy, therefore, thousands, training, transferred, types, unable, vocabulary, web, whenever

Student 40

Abilities, accident, according, advantages, aggressive, aid, airline, appeal, assignment, attract, blame, bored, boss, career, chair, communication, company, conclusion, conduct, confidence, conflict, contract, considerate, customers, deal, decrease, delay, domain, drugs, earthquake, economic, enabled, encourage, endangered, essay, extinct, first, finally, flexibility, flood, future, gender, globalization, hire, homes, homework, hundred, improved, individuals, intelligent, international, legal, lifestyles, mental, mention, ministry, moreover, nevertheless, nowadays, opportunities, past, physically, pivotal, population, properties, rates, reduce, responsibility, salary, sector, senior, social, spill, solve, support, sympathy, tanker, therefore, third, tourism, view

Student 41

According, affect, aid, avoid, basic, benefit, communicate, companies, conclusion, consumer, culture, decrease, dramatically, economics, encouraging, endangered, essay, extinct, facilities, finally, firstly, flexible, flight, globalization, graduation, healthcare, homework, hire, hundred, improving, informal, major, million, moreover, mutual, negative, niece, nowadays, origin, percent, population, preventing, projects, pupils, rates, receive, reduce, region, replaced, responsibilities, restrictions, returned, robot, salary, ships, social, solution, stereotypes, supporting, technology, third, thousand, types, whenever

Student 42

Academic, according, achieve, advantages, airport, appropriate, arrived, assessment, balance, brainstorming, center, coast, college, company, comprehension, conclusion, conflict, context, consumer, coral, county, cultural, dealing, desert, diamond, dialogue, disadvantages, donkey, dynamite, economically, enforce, entrance, evaluate, essay, explorer, favorite, feather, finally, flexible, future, globalization, goat, graduate, gym, heart, hit, homes, homemade, huge, hundred, identifying, income, million, model, monsoon, moreover, nervous, neutral, ninety, nowadays, opinion, permanently, professional, promotion, protecting, rate, reef, recording, responsibility, returned, salary, sheep, shrub, shift, similarity, socially, solve, solution, spill, standardized, stakes, structure, support, surround, tankers, third, training, trip, undergraduate, variety, wonderful

Student 43

Ability, able, according, adapt, admired, aid, amazing, arrives, average, barriers, beach, beef, benefits, biology, breakfast, cafeteria, center, challenge, chess, combine, companies, computer, conclusion, consequently, crime, culture, dairy, deal, decrease, design, determine, dinner, disaster, discrimination, economic, endanger, enter, especially, essay, exchange, extinct, flashlight, finally, first, furthermore, future, globalization, goal, graduate, guide, gulf, hire, hybrid, impact, immigrate, influenced, juggling, kit, legal, lunch, mall, memory, museum, native, negative, nineteen, numerous physics, objective, outlined, outsourcing, past, project, protect, population, pupils, rates, reduce, replacement, responsibility, retires, salary, section, sector, shocked, shower, social, solutions, solve, specific, split-shift, stake, standardized, strategies, tent, third, thirty, traditional, training, trip, type, wonderful

Student 44

Able, according, activities, advanced, advantage, areas, balance, blame, borders, citizen, college, communication, company, computer, conclude, conclusion, conflict, considerable, culture, decrease, destruction, discrimination, encourage, endangered, enforce, environment, essay, extinct, express, facilities, favorite, field, finally, firstly, flexible, focus, gains, globalization, graduated, gulf, habitat, heart, hiring, hundred, huge, ignore, immigrate, internet, item, located, major, marine, ministry, moreover, offering, opportunities, outsourcing, overseas, oxygen, past, percent, phone, population, projects, promotion, protect, rate, reduce, reefs, remain, salary, security, solution, solve, spreads, strategy, strict, summarize, support, tanker, telecommunication, third, thousand, unless, variations, view

Appendix 11
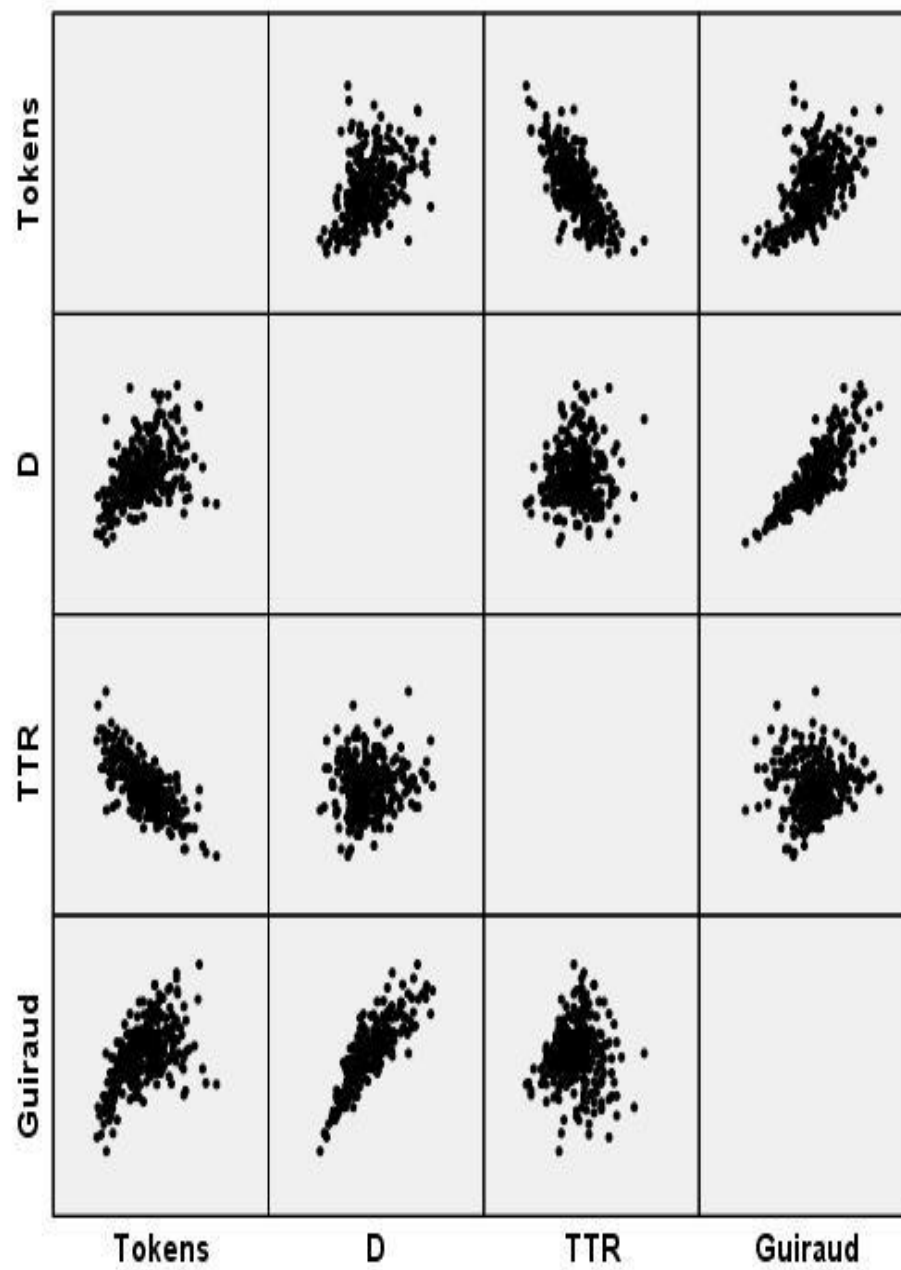
Appendix 12

Scatterplot for correlation between measures of lexical sophistication
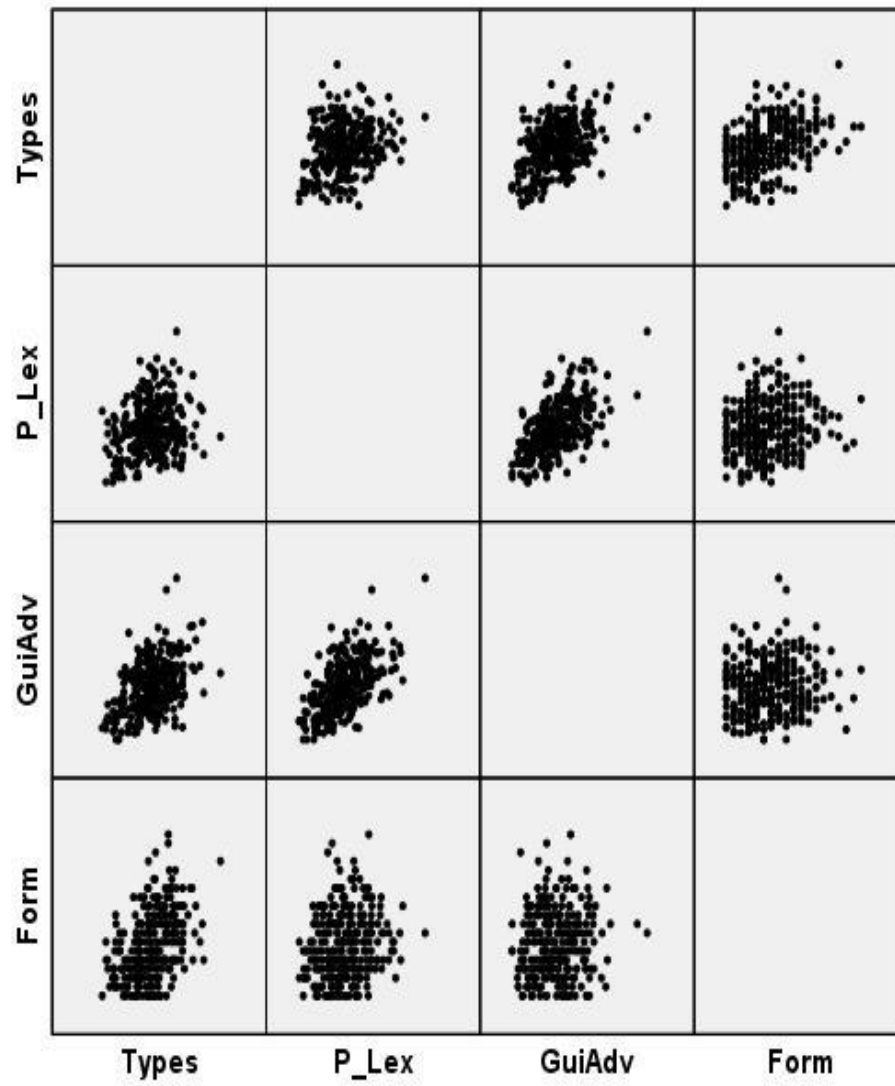
Appendix 13

<u>Grade 8 Essays</u>

Essay 10.4

Total number of formulaic sequences: 7

there are (2)

based on (1)

instead of (2)

all over (1)

out of (1)

Essay 8.4

Total number of formulaic sequences: 2

there are (1)

cause (1)

<u>Grade 7 Essays</u>

Essay 2.6

Total number of formulaic sequences: 9

there are (3)

carry out (1)

deal with (2)

each other (1)

lead to (1)

come up with (1)

Essay 28.6

Total number of formulaic sequences: 8

as well as (1)

is to (1)

such as (3)

even though (1)

in full (1)

by way of (1)

Essay 28.7

Total number of formulaic sequences: 15

a lot (1)

a number of (1)

as well (1)

lead to (2)

tend to (1)

even though (1)

that is (1)

get to (2)

one another (1)

a good (1)

contrary to (1)

over time (1)

can tell (1)

Essay 23.5

Total number of formulaic sequences: 12

a lot (1)

going to (1)

have to (1)

is to (1)

there are (3)

deal with (1)

look for (1)

instead of (1)

all over (1)

in my opinion (1)

Essay 25.5

Total number of formulaic sequences: 8

have to (1)

there are (1)

think about (1)

at the same time (1)

take care of (2)

at work (1)

cause (1)

Grade 6 Essays

Essay 38.5

Total number of formulaic sequences: 8

have to (1)

there are (1)

there is (1)

too much (1)

for all (1)

at work (1)

cause (2)

Essay 18.2

Total number of formulaic sequences: 6

there are (1)

kind of (1)

put it (1)

for all (1)

cause (2)

Essay 18.3

Total number of formulaic sequences: 12

going to (2)

there are (5)

instead of (1)

cause (4)

Essay 21.3

Total number of formulaic sequences: 4

kind of (1)

to go (2)

more and more (1)

Essay 37.3

Total number of formulaic sequences: 7

have to (1)

there are (2)

cause (4)

Essay 37.6

Total number of formulaic sequences: 12

I mean (2)

there are (1)

there is (1)

deal with (1)

that is (2)

at work (3)

kind of (2)

Essay 44.4

Total number of formulaic sequences: 10

a little (1)

a lot (1)

there are (2)

focus on (1)

take advantage (1)

a good (1)

cause (3)

Essay 4.6

Total number of formulaic sequences: 10

as well as (1)

going to (1)

there are (2)

in addition (1)

too much (1)

think it is (1)

a good (2)

in my opinion (1)

Essay 27.6

Total number of formulaic sequences: 10

each other (1)

find it (1)

at work (8)

Essay 36.5

Total number of formulaic sequences: 8

going to (1)

there are (3)

deal with (1)

in addition (1)

in the end (1)

take care of (1)

Essay 44.5

Total number of formulaic sequences: 8

have to (2)

there are (2)

work on (1)

focus on (1)

take care of (2)

Essay 19.7

Total number of formulaic sequences: 6

is to (1)

such as (1)

there are (1)

used to (1)

out of (1)

used to (1)

Essay 25.7

Total number of formulaic sequences: 5

there are (1)

all over (1)

a good (3)

Essay 36.7

Total number of formulaic sequences: 7

in fact (1)

so that (1)

such as (1)

there are (1)

think about (1)

a good (1)

provide for (1)

Essay 43.7

Total number of formulaic sequences: 4

there is (2)

instead of (1)

cause (1)

Essay 38.3

Total number of formulaic sequences: 4

there are (2)

cause (2)

Essay 17.4

Total number of formulaic sequences: 7

have to (1)

such as (2)

there are (2)

cause (2)

Essay 21.6

Total number of formulaic sequences: 9

going to (2)

such as (1)

there are (2)

for all (1)

to go (1)

at work (1)

take care of (1)

Essay 36.5

Total number of formulaic sequences: 8

going to (1)

there are (3)

deal with (1)

in addition (1)

in the end (1)

take care of (1)

Essay 29.6

Total number of formulaic sequences: 8

there are (2)

there is (1)

deal with (4)

good at (1)

Essay 29.7

Total number of formulaic sequences: 10

deal with (3)

each other (2)

you see (1)

that is (1)

kind of (1)

to me (1)

or anything (1)

<u>Grade 4 Essay</u>

Essay 27.2

Total number of formulaic sequences: 17

a lot (6)

of course (3)

there are (5)

kind of (1)

for all (1)

cause (1)

<u>Grade 2 Essay</u>

Essay 34.1

Total number of formulaic sequences: 1

there are (1)

Appendix 14

IELTS SAMPLE ESSAYS -TRANSCRIPTIONS

Essay 1AA – Band 5

@Begin

*TXT: This is a bar chart of the number of men and women in further education in Britain in three periods.

*TXT: In 1970, most of men were studying part-time but from 1980, studying part-time was decreased and studying full-time was increased and in 1990, it was twice as many students as 1970.

*TXT: On the other hand, women studying full-time were increased and not only full-time, part-time also were increased.

*TXT: In 1990, studying full-time was three times as many students as in 1970.

*TXT: If compare men and women, as you see, in 1970, men were studying more than women full-time or part-time but it changed from 1980 and then.

*TXT: In 1990, women were studying part-time more than men and studying full-time was same number.

*TXT: It shows you women has a high education now.

@End

Essay 1AB- Band 6

@Begin

*TXT: According to this graph, the number of men and women in further education in Britain shows the following patterns.

*TXT: In the case of male, the number of male has declined slightly from about 1000 thousands in 1970/71 from about 820 thousands in 1980/81.

*TXT: However, this figure rose back to about 850 thousands in 1990/91 from about 820 thousands in 1980/81.

*TXT: The proportion of full-time education has declined during this period.

*TXT: However, the proportion of part-time education has increased dramatically.

*TXT: On the other hand, in the case of female, the number of both full-time education and part-time education has increased during this period.

*TXT: From about 700 thousands in 1970/71, these figures rose to about 820 thousands in 1980/81, to about 1100 thousands in 1990/91.

*TXT: In terms of full-time education, this figure rose by about 260 to about 900 in 1990/91.

*TXT: On the other hand, with respect to part-time education, this figure rose dramatically between 1980/81 and 1970/71.

*TXT: However this figure rose slightly between 1980/81 and 1990/91. @End

Essay 1BA- Band 6

@Begin

*TXT: The graph shows the percentage of audiences over 4 years old of UK follows the radio and television throughout the day during the period October December 1992.

*TXT: It has been observed from the graph that less than 10% audiences follows the radio at 6:00 am and the percentage raised to a pick around 30% at 8 am and decline gradually to around 10% during the period 2:00 to 4:00 pm and again raised a bit to around 12% between 4:00 to 6:00 pm then again dropped to below 10% at around 10 pm.

*TXT: The rate again raised to a bit between 10:00 pm to 12:00 pm and then dropped slowly by 4:00 am.

*TXT: On the other hand, the rate of television audiences raise 0-10% during the period 6:00 to 8:00 am and remain steady up to 10 am and then gradually goes down by 12:00 am.

*TXT: The percentage raised dramatically to around 10% by 2:00 pm which again raised to a pick above 40% between 6:00-8:00pm and then gradually dropped between the period 12:00 pm to 4:00 am.

@End

Essay 1BB- Band 7

@Begin

*TXT: The bold graph shows the television audiences throughout the day.

*TXT: It shows that the percentage of audiences is zero percent in early morning but it gradually rises up to ten percent at 8:00 am and maintains the same for the next two hours.

*TXT: There is a slight fall in percentage in next two hours however after that it rises sharp up to twenty percent within the next two hours.

*TXT: After this the graph rises very fast and attains its' peak at 10 pm which is about forty five percent.

*TXT: The graph gradually falls down and at 2:00 am it is at five percent.

*TXT: The thinner graph shows the percentage of radio audiences.

*TXT: Unlike the television one the peak percentage of the radio audiences is at 8:00 am which is about 30 percent.

*TXT: Then it gradually falls and it corresponds with the television one at two pm.

*TXT: After that it gradually falls but with a small increase in percentage at 4:30 to 6:00 pm.

*TXT: The percentage of audience then gradually goes down and at four am it is the lowest which is near 2 percentage.

*TXT: These graphs prove the progressive popularity of television.

@End

Essay 2AA- Band 5

@Begin

*TXT: Nowaday, there are a lot of cars on British road and they have increased day to day.

*TXT: By the year 2000 there may be as many as 29 million vehicles on British roads.

*TXT: In this essay, I intend to examine, about the solutions of these problems.

*TXT: Firstly, the people living in Britain need to think about themselves.

*TXT: If they used the bus and train instead of their car, this problem would resolve a little.

*TXT: Because of this, the British government should introduce to control car ownership and use.

*TXT: For example, the government can ban to enter the road by car in the same day all family from a house.

*TXT: Secondly, the buses and trains of government should be free for public population.

*TXT: Thus, the people would use these transport vehicles instead of their own car.

*TXT: After that, the roads in Britain would be safer and more comfortable.

*TXT: Lastly, the number of cars that are exported from another country should decrease and the prices of car should increase in case they are not overcrowded.

*TXT: For example, the prices of cigarettes increased and the consumption of cigarettes went down.

*TXT: In conclusion, if these measures put into action the problem of traffic can be decreased in the British roads.

@End

Essay 2AB- Band 6

@Begin

*TXT: The transport has been one of the most important problems for the last two centuries.

*TXT: The problem began with the development and the growing of the cities.

*TXT: Before the eighth century the people lived in small villages or towns and did not have necessity to go too far.

*TXT: The people did not worry about the time to arrive in some where.

*TXT: Nowadays the situation changed.

*TXT: Many cars on the streets and many people need to go to any place.

*TXT: The numbers of cars has increased and as a result there are many problems: pollution, noise, car accident, insufficient car park and petroleum problem.

*TXT: On the other hand, people use car to go anywhere: to work, to travel, to spent holiday and to amusement.

*TXT: Meanwhile the car is important the cities must have another solution.

*TXT: It is important to organise its using and to meet alternative ways.

*TXT: In big cities there are some alternatives like undergrounds (metro), coach, train and bicycles.

*TXT: In China and Cuba for example they use a lot of bicycles for substituting the cars and coaches.

*TXT: It would be better to think about others differents kinds of transport.

*TXT: In Brazil the government has talked about transport on the rivers.

*TXT: In this country there are many rivers where it is possible to go to different places.

*TXT: In general they are flat rivers.

*TXT: Another kind of transport is car that uses solar energy.

*TXT: Probably they don't have pollution problem and it is cheaper than others car.

*TXT: In conclusion, the transport is a social problem in big cities but its solution depend on new technologies, others kind of energy and political aspects.

@End

Essay 2BA- Band 5

@Begin

*TXT: Nuclear power provides cheap energy sources.

*TXT: Sometimes the present sources and energy like oil, gas etc. will be finished.

*TXT: Arguments in favour nuclear power: the nuclear energy produces by chemical materials: it is comparatively cheaper than other energy.

*TXT: To produce the power it only involve some expert people and energy plant.

*TXT: Where to produce other energy it needs large involvement like worker, machineries, etc.

*TXT: And also takes more time.

*TXT: The nuclear power plants are well-protected and monitor.

*TXT: That is why there is less possibility.

*TXT: The threat of nuclear weapons maintains world peace because the developed countries like UK, USA, Canada, France etc. have similar weapons (warhead).

*TXT: Each country do not give threat to the country.

*TXT: Because they know if the country destroys cities, then other will create problems from them.

*TXT: So it is well-balanced and world peace maintains peacefully.

*TXT: Though there are sometimes creates problems by the nuclear technology but sometimes it also help the mankind in the field of medical and engineering sectors.

*TXT: In the medical field we can say by nuclear way sometimes we can treat a cancer patient.

*TXT: On the other hand in the field of engineering by the nuclear power engineers can do lot of things like operate engine instead of electricity.

*TXT: In conclusion we can say though there are some problems in the nuclear power but it lies some benefit for the mankind.

@End

Essay 2BB- Band 7

@Begin

*TXT: Nuclear power is an alternative source of energy which is carefully being evaluated during these times of energy problems.

*TXT: During these years we can say that we have energy problems but in more or less 50 years, we will be facing an energy crisis.

*TXT: Nuclear power is an alternative source of energy and unlike other sources such as solar energy, nuclear power is highly effective for industrial purposes.

*TXT: If it is handled correctly there really is no danger for the public.

*TXT: It is cheap, there is no threat of pollution and best of all it is limitless.

*TXT: It is difficult to think about nuclear power as a good source of energy for people in general.

*TXT: This is due to the use it has been given since its birth during the second world war.

*TXT: It is expressed as military power and in fact at the moment nuclear power is limited to few hands who consider themselves world power.

*TXT: When and if there is a change of ideology regarding the correct use of nuclear power, then we may all benefit from all the advantages nuclear power can give us.

*TXT: If we outweigh the advantages and disadvantages of nuclear technology we often have the following: As stated before, the advantages are that there is limitless supply, it is cheap, it is effective for industrial purpose and still there are many benefits which have not yet been discovered.

*TXT: The disadvantages are at present time that it is limited to only a few countries who regard it as safe military power.

*TXT: Also if mishandled, there is risk for the population around the plant to undergo contamination as we all happened in Chernobyl.

*TXT: If these disadvantages can be overcome, then it is clear that nuclear energy can give us more benefits than problems.

*TXT: It will in the future be very important as the energy crisis is not far ahead.

*TXT: In conclusion, nuclear power is good, it can be safe, and we will all benefit.

*TXT: It is up to our leaders to see that it is handled well so that we can all benefit from it.

@End

Appendix 15

IELTS SAMPLE ESSAYS- EXAMINERS' COMMENTS

Essay 1AA

Examiner comment
Band 5
The length of the answer is just acceptable. There is a good attempt to describe the overall trends but the content would have been greatly improved if the candidate had included some reference to the figures given on the graph. Without these, the reader is lacking some important information. The answer is quite difficult to follow and there are some punctuation errors that cause confusion. The structures are fairly simple and efforts to produce more complex sentences are not successful.

Essay 1AB

Examiner comment
Band 6
The candidate has made a good attempt to describe the graphs looking at global trends and more detailed figures. There is, however, some information missing and the information is inaccurate in minor areas. The answer flows quite smoothly although connectives are overused or inappropriate, and some of the points do not link up well. The grammatical accuracy is quite good and the language used to describe the trends is wellhandled. However, there are problems with expression and the appropriate choice of words and whilst there is good structural control, the complexity and variation in the sentences are limited.

Essay 1BA

Examiner comment
Band 6
The answer has an appropriate introduction which the candidate has attempted to express in his/her own words. There is good coverage of the data and a brief reference to contrasting trends. The answer can be followed although it is rather repetitive and cohesive devices are overused. In order to gain a higher mark for content, the candidate would be expected to select the salient features of the graph and comment primarily on these. Sentences are long but lack complexity. There are some errors in tense, verb form and spelling which interfere slightly with the flow of the answer.

Essay 1BB

Examiner comment

Band 7

The answer deals well with both the individual media trends and the overall comparison of these trends. The opening could be more fully developed with the inclusion of information relating to the groups studied and the period of time during which the study took place. There is a good variety of cohesive devices and the message can be followed quite easily although the expression is sometimes a little clumsy. Structures are complex and vocabulary is varied but there are errors in word forms, tense and voice though these do not impede communication.

Essay 2AA

Examiner comment
Band 5

The answer is short at just over 200 words and thus loses marks for content. There are some relevant arguments but these are not very well developed and become unclear in places. The organisation of the answer is evident through the use of fairly simple connectives but there are problems for the reader in that there are many missing words and word order is often incorrect. The structures are quite ambitious but often faulty and vocabulary is kept quite simple.

Essay 2AB

Examiner comment
Band 6

There are quite a lot of ideas and while some of these are supported better than others, there is an overall coherence to the answer. The introduction is perhaps slightly long and more time could have been devoted to answering the question. The answer is fairly easy to follow and there is good punctuation. Organisational devices are evident although some areas of the answer become unclear and would benefit from more accurate use of connectives. There are some errors in the structures but there is also evidence of the production of complex sentence forms. Grammatical errors interfere slightly with comprehension.

Essay 2BA

Examiner comment
Band 5

Although the script contains some good arguments, these are presented using poor structures and the answer is not very coherent. The candidate has a clear point of view but not all the supporting arguments are linked together well and sometimes ideas are left unfinished. There is quite a lot of relevant vocabulary but this is not used skilfully and sentences often have words missing or lapse into different styles. The answer is spoilt by grammatical errors and poor expression.

Essay 2BB

Examiner comment
Band 7
The answer is wellwritten and contains some good arguments. It does tend to repeat these arguments but the writer's point of view remains clear throughout. The message is easy to follow and ideas are arranged well with good use of cohesive devices. There are minor problems with coherence and at times the expression is clumsy and imprecise. There is a wide range of structures that are well handled with only small problems in the use of vocabulary, mainly in the areas of spelling and word choice.