The Role of Lexical Collocations and Learner and Course Variables in Determining Writing Quality in Assignments from a First Year Composition Programme

Thesis Submitted by

Lee McCallum

to

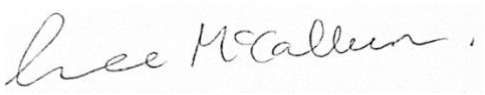
The University of Exeter

for the degree of

Doctor of Education in TESOL

Submission date

June 2021

(Signature) 

Abstract

Over the last decade, studies have shown how the diverse and sophisticated use of collocations in student writing has a relationship with grades awarded by raters. However, these studies have a number of contextual and methodological limitations. Contextually, studies have largely been limited to second language writers whose writing has been evaluated according to the Common European Framework of Reference (CEFR). Methodologically, these studies have used a small number of linguistic measures of sophistication in the form of association measures. Such a limitation is a result of the guiding measurement literature appearing in a fragmented manner. This fragmented manner does not make bringing together the range of possible measures we may use an easy task. Further still, the relationship between collocations and writing quality has been studied using correlation and/or multiple linear regression techniques. These techniques do not consider the effect that the hierarchical nature of the corpus may have on measuring this relationship. This means previous studies have not accounted for the random effect of individual writers or raters in their analyses.

The present study tackles these limitations. The study adds contextual weight to the literature by investigating the relationships between collocations and writing quality in a corpus of First Year Composition (FYC) essays written by speakers of English as a first and second language in the US. To tackle the methodological limitations, the research is divided into two studies. The first study examines the theoretical and mathematical connections between a subset of association measures via the use of a cluster analysis.

From this cluster analysis a case is made as to which measures should be selected for inclusion in the subsequent investigation of the collocation-writing quality relationship. The second study also tackles methodological limitations by modelling the relationship between collocations and writing quality through a mixed effects logistic regression model which accounts for random writer and rater effects. Considered individually, each study has implications for how we measure the relationship between collocations and writing quality. Study one has implications for discussions around measure use and how we understand the properties of collocation. Study two has implications for understanding the relationship collocations have with the grades writers are awarded on the FYC programme. Taken together both studies allow researchers to consider how such relationships may be measured in the future so as to appreciate a broader range of contexts, appropriate objective measure selection and rigorous statistical techniques.

Keywords: Collocation, Writing Quality, Mixed Effects Models, FYC.

Dedication

To my parents

Acknowledgements

Few people understand the journey involved in doctoral study in terms of the countless years of reading, writing and being alone with a research problem that threatens to overcome you. However, I am extremely fortunate to have had the support of several people along the way. Firstly, this study would not have been possible without the constant support of my supervisor: Dr. Philip Durrant. From taking me under his wing to write articles and a book to answering my constant emails, he went above and beyond what a supervisor should be. I don't know what to say except: thank you for everything, Phil. Neither words nor actions will ever be able to repay you! The study also owes a great deal to my second supervisor: Dr. Dongbo Zhang whose enthusiasm and knowledge added immensely valuable input to the research.

Huge thanks must go to the institution in the USA where this study is based. Their generosity in providing access to texts, faculty and bringing on board members of their team who could add value and expertise to the study made the study not only possible but as strong as possible within its respective time limits. In particular, I would like to thank Dr. Joseph Moxley, Dr. Norbert Elliott, Dr. Nicole Tracy-Ventura and Dr. Lisa Meloncon who went above board to meet my needs as an external researcher whilst keeping me well informed in what I needed to do to comply with the university's requirements. The study also benefited immensely from insights gained from the faculty involved directly in the FYC programme namely: Dr. Dianne Donnelly, Dr. Alaina Tackitt and the several instructors who provided comments on draft chapters. I hope the final study is of value to you and your team of dedicated professionals.

I would also like to thank a number of first and second language academics who supported me along the way. Thanks go to Ute Römer at Georgia State University who kindly sent me MICUSP texts when I asked for them. Without your help, Ute, I would have spent many more hours on the preparation of a reference corpus. Thanks also go to Christine Coombe, Mark Brenchley and Niall Curry who will never know how much I appreciated their emails that checked up on me and my progress. Thanks for listening!.

Thanks, must also go to my many pals around the world who have played a role in my career and kept me sane throughout the whole lonely journey that a doctorate can be. A special mention goes to my dear friend Romyna, who never stopped encouraging me. Thank you so much.

On a personal level, I owe a great deal to my family especially my long-suffering husband Mauro. We've come a long way from living apart for 3 years and moving around Scotland to save money and seek out job opportunities. I couldn't have done this without you.

To my parents, from the day I was born weighing less than a kilo, you've never left my side and always believed in me. Neither of you know anything about university or research but never stopped listening when I needed you to. Thank you for putting up with me for the last 30 years and for putting me up when I taught around Scotland and needed a bed for several nights. I will one day repay the rent money!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AP: Advanced Placement
BNC: British National Corpus
CEFR: Common European Framework of Reference
COCA: Corpus of Contemporary American English
EFL: English as a Foreign Language
ENC 1101: First FYC module
ENC 1102: Second FYC module
ESL: English as a Second Language
FCE: First Certificate of English
FYC: First Year Composition
GTAs: Graduate Teaching Assistants
ICLE: International Corpus of Learner English
IELTS: International English Testing Service
L1: First language
L2: Second language
My R: My Reviewers
NLP: Natural Language Processing
POS: Part of Speech
TAALES: Tool for the Automatic Analysis for Lexical Sophistication
TOEFL: Test of English as a Foreign Language
TTR: Type Token Ratio
USF: University of South Florida
WPA: Writing Programme Administration
WPAs: Writing Programme Administrators

Chapter One: Introduction

1.1 Introduction

This study is about measuring the relationship between the use of collocations and writing quality grades in a set of student essays from a First Year Composition (FYC) university programme in the US. The study concerns why and how we measure this relationship and the implications of such measurement for scholars in the linguistic feature – writing quality community. The study also informs the wider conversations that are taking place around providing more language focused instruction on FYC programmes.

The first chapter to the study introduces the rationale for a focus on the linguistic area of collocation and sets out the case for such a focus to take place in a First Year Composition (FYC) context in the US. In doing so, the chapter explains the background to the research and its key constructs (Section 1.2) and the explicit limitations that the study addresses (Section 1.3). The chapter then outlines the specific contributions that can be made by addressing these limitations and the broad aims of such work (Section 1.4), the study's research questions (Section 1.5) and the overall organisation of the study (Section 1.6).

1.2 The Background to the Study

The relationship between linguistic features and writing quality[1] has been an active area of scholarship across L1 and L2 educational contexts for several decades (e.g., Arthur, 1979; Hunt, 1970; Kameen, 1979). Many scholars have tapped into this relationship by relying on qualitative accounts from instructors and students that explain how linguistic features are used in written texts (e.g., Hirose & Sasaki, 1994; Sasaki & Hirose, 1996). However, in the last two decades, most of the work that has informed understandings of this relationship has increasingly relied on quantitative corpus and computational-based methodologies that aim to measure this relationship on a largely straightforward objective basis (e.g., Berman & Nir-Savig, 2007; Berman & Nir, 2010; Crossley & McNamara, 2012; Cumming et al., 2005; Durrant, Brenchley & McCallum, 2021; Eckstein & Ferris, 2018; Grant & Ginther, 2000; Kim, Crossley & Kyle, 2018; Kyle & Crossley, 2016; Paquot, 2018, 2019; Staples & Reppen, 2016).

Broadly speaking, most of this research has had the aim of measuring the strength of this relationship statistically via correlation and multiple linear regression analyses. These analyses have informed classroom pedagogy and rating scale development by identifying pertinent features and incorporating them into teaching and rating scale descriptors (e.g., Biber & Gray, 2013; Hawkins & Filipovic, 2012). At the same time, they have also been used to train and modify large-scale written feedback and grading software by illuminating the linguistic features that *appear* to have a relationship with human judgements of writing quality (e.g., Attali & Burstein, 2006 & Deane & Quinlan, 2010). Given these applications to

[1] Throughout this study, the term writing quality will be used synonymously with that of writing proficiency. The two terms are used interchangeably to symbolise 'better' or more proficient writing. A discussion of this construct for the FYC context takes place in Chapter Two (Section 2.5).

practice, there is an assumption then that the identified features have a relationship to the overall programme, exam or assignment success that students achieve; and researchers are able to use these features to predict this success.

These features have been primarily studied under three linguistic areas of interest. These are *grammatical* features that operate at clause or phrasal level (e.g., number of clause types, number of finite clauses) (e.g., Biber et al., 2011; Bulté & Housen, 2014; Crossley, Kyle & McNamara, 2016; Ferris, 1994; Hou, Verspoor & Loerts, 2016; Taguchi, Crawford & Wetzel, 2013; Zheng, 2016); *individual vocabulary items* (e.g., number of lexical types, % of words used that appear in the Academic Word List) (e.g., Daller, Turlik & Weir, 2013; Horst & Collins, 2006; Laufer & Nation, 1995; Eckstein & Ferris, 2018); and *phraseology* with increasing attention paid to the notion of collocation (e.g., Bestgen & Granger, 2014; Granger & Bestgen, 2014; Paquot, 2018, 2019; Kyle & Crossley, 2016; Garner, 2018, 2019).

The sections that follow outline what collocations are, how they have been identified and why they are perceived as an important feature of academic writing.

1.2.1 Properties of Collocation and their Importance in Academic Writing

The ability to write well is seen as a hallmark of not only language proficiency but also one of academic success for L1 and L2 students (Richards & Renandya, 2002, cited in Nosratinia & Razavi, 2016; Tang, 2012). At university, the importance of writing is evident as it makes up a large proportion of academic coursework and end of semester or year examinations (Hyland, 2013; Zhu, 2004). This importance often runs alongside the status writing has as the most difficult skill for L1 and L2 students to master given the repertoire of language items it involves, and the level of knowledge needed to link them together

(Crossley et al., 2011; Llach, 2007). This means writing is seen as a difficult skill which must be mastered if students want to achieve academic success. In the U.S, this struggle is recognised in that both L1 and L2 student populations are often required to enrol in FYC writing programmes at university. Enrolment on these writing programmes allows students to develop the necessary composition skills that will prepare them for the academic writing that they will encounter as they progress through their studies (Aull, 2015a; Connors, 1997).

In becoming familiar with university writing, both L1 and L2 students are expected to adopt academic language that is befitting of the genres to which their FYC programmes attempt to introduce them to. In this respect, use of this language is taken to be a key marker of achieving success and meeting FYC programme learning outcomes (Framework for Success in Postsecondary Writing, 2011; CWPA Outcomes Statement, 2014). However, adopting this language use is not without its challenges for both student groups as they attempt to 'fit into' their respective discourse communities (Aull, 2015a, 2015b, 2015c,2017,2019; Bartholomae, 1986; Wray, 2002). In order to fit into their communities, students are expected to write and communicate in language that is expected of that community. Wray (2006, p. 593) sums this up by noting that: "when we speak, we select particular turns of phrase that we perceive to be associated with certain values, styles and groups". The learning of these specific word combinations is seen as a "badge of identity". This "badge of identity" involves fitting into a particular discourse community and learning to become a proficient writer who is able to navigate the language expected in such a community; and with that is said to be concerned with the mastery of using word combinations, especially those termed collocations (e.g., See Durrant, 2019).

The notion of collocation, word combination or partnership operates under the diverse umbrella term of 'formulaic language'. Drawing on Wray's (2002, p.9) definition of a formulaic sequence as: " A sequence, continuous or discontinuous, of words, or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" emphasises that formulaic language concerns word strings or stretches of words that are thought of as single items. In a later definition, Wray (2019, p. 267) comes back to this notion of a formulaic sequence being a single lexical unit by stating that a formulaic sequence is: "any multiword string that is perceived by the agent (i.e. learner, researcher, etc.) to have an identity or usefulness as a single lexical unit". Siyanova-Chanturia and Pellicer-Sánchez (2019, p.6) unravel the notions of identity or usefulness by indicating that the identity or use is hemmed in by the interests and views of the agent or the person who decides on how a particular lexical unit should be perceived:

> It may entail high frequency of occurrence (since frequently produced strings, other than being useful by virtue of being frequent in language, may also benefit from being treated as a single unit), a teacher's perceived value of a string (no matter how frequent), some sort of basic holistic storage and processing, a specific pragmatic function, or, indeed, something altogether different (Siyanova-Chanturia & Pellicer-Sánchez, 2019, p. 6).

Operating within this broad umbrella term of formulaic language, collocations are considered to be hazily defined word combinations which vary in length (combinations can be stretches of language that are typically two to nine words in length) and may or may not be contiguous combinations (i.e. appearing in uninterrupted sequence). More widely drawn on when defining collocations are two other features: their frequent nature in language and the extent that the individual words can be substituted for others in the combination. Collocations are defined as such when the combination of words appears more frequently

than the frequencies of the individual words would be able to predict (Jones & Sinclair, 1974). A second defining feature is that in these word combinations there is a degree of transparency in meaning and substitutability where one or more of the words in the combination can be substituted for another. Often this substitutability is arbitrary. These defining features of frequency and substitutability are often referred to as two different paradigms that shape the definition, extraction and study of collocations. These paradigms are the frequency-based approach and the phraseological approach. In the former, the frequency of the word combination offers an insight into its collocational nature; while in the latter the semantic properties of transparency and substitutability are considered more important. Drawing on these paradigms, the literature (e.g., Granger & Paquot, 2008; Manning & Schütze, 1999; Paquot, 2018, 2019) has referred to examples (1) – (5) as collocations:

(1) Apologise profusely

(2) Strong tea

(3) Weapons of mass destruction

(4) Heavy rain

(5) Rancid butter

An important part of our understanding of these word combinations lies in the fact that the individual words in examples (1) – (5) *could* be analysed individually as smaller parts but if or when we chose to do so, we lose valuable linguistic meaning. These combinations are learned as wholes because the learning of the individual words, does not facilitate the understanding or ability to use the combination in its entirety (Choueka 1988; Evert, 2004; Nesselhauf, 2005; Durrant et al., 2021). For example, learners may know the individual

words '*on the ball'*, but they are unlikely to produce '*on the ball*' as a collocation unless the combination and its meaning have been specifically learned (Durrant et al., 2021). These word combinations are treated holistically as single units by learners and teachers whereby the combinations represent a single choice for language users in their lexicons (Palmer, 1933/1966; Sinclair, 1987). In this sense then, examples (1) – (5) are learned as single units and not a series of individual words. Importantly, the holistic single unit principle means that researchers also identify, extract, and analyse these word combinations as single units. In doing so, researchers need to delimit word combinations that are taken to be collocational from those which are thought to be non-collocational.

Through a frequency-based lens, this distinction between collocation and non-collocational word combinations has been determined through the use of mathematical formulas. These formulas are known as measures of association, whereby their formulas are able to distinguish whether or not the words in the combination occur together by chance or not (Evert, 2004) (these formulas are set out fully in Chapter Four, Sections 4.2 – 4.4). In contrast, through the phraseological approach, collocations are identified as word combinations which take on particular meanings (e.g., '*curry favour*') that they do not have in other environments, or are determined by what extent their make-up is restricted arbitrarily from a semantics-driven perspective (e.g., the combination of the verb '*commit*' with doing something wrong or illegal) (Gablasova et al., 2017; Durrant & Schmitt, 2009). Nesselhauf (2005) points out that under the phraseology-based perspective, there is a greater emphasis on the degree that one word in the combination is fixed according to semantic preferences i.e. to what extent words can be substituted for others in the combination, and to what extent meaning is literal and transparent.

In this study, the notion of collocation is underpinned by the frequency-based perspective and narrows to use the term '*statistical collocation*' which aims to determine a collocation from its frequency profile or statistical make up. The term statistical collocation can be traced back to Jones and Sinclair (1974, p.19) who state that statistical collocations are discontinuous word pairs that "co-occur more often than their respective frequencies and the length of text in which they appear would predict". Schmitt and Schmitt (2020, p.5) also comment that word partnerships or collocations are governed by two crucial factors: the idea that words co-occur together and that these co-occurrences have varying degrees of exclusivity. They exemplify these factors through a narrative about the word '*blonde*'. They note that '*blonde*' occurs almost exclusively with the word '*hair*' or a few other nouns such as '*woman*' or '*lady*' noting that we say and write '*blonde hair*', '*blonde woman*' or '*blonde lady*' but never '*blonde wallpaper*' or '*blonde paint*' (italics in authors' original), although these latter combinations are syntactically and semantically possible. They further add that this notion of partnership varies with some words collocating strongly and others less so. They further exemplify co-occurrence and exclusivity through the word '*nice*' which commonly occurs with words that we associate with pleasantness, for example, '*nice view*', '*nice car*', '*nice salary*'; with these examples collocating less so than the examples given for '*blonde*' and '*hair*'. With respect to understanding the connection between co-occurrence and exclusivity, they use the example of '*the*' which co-occurs or appears alongside almost every non-proper noun and therefore yields little evidence to warrant the notion of collocation. They summarise then that to be considered a collocation, the words must co-occur but also have an element of exclusivity.

Considering this, the study's framing of collocation is therefore guided by the fact that in these word combinations, one word is used *almost exclusively* in combination with another word. Clear (1993) makes this apparent in his analysis of '*arbiter taste*'; while examples from Durrant and Schmitt (2009) include '*tectonic plates*', '*global warming*' and '*ethnic minorities*' where we see that word one has few, if any, alternative partners. This connection to frequency becomes clear when we consider the views of Firth (1968), for whom, collocation is a type of mutual expectancy between words – where see one, we expect to see the other. The words are therefore said to predict each other. The linguistic examples [1] – [5] also help reinforce the concept of exclusivity. In these examples, few things can be rancid (other than butter), few things can be done profusely (other than the act of apologising); yet it is also plausible that more nouns can be described as strong and/or heavy: reinforcing the idea that some words partner others in an exclusive capacity, that is to say that they have few, if any alternative partners, while others may have many more partners and therefore be considered less exclusive and more generic in use. These examples highlight that exclusivity as a property of collocation operates not absolutely but on a sliding infinite scale whereby word combinations can be compared.

1.2.2 The Importance of Collocation to Student Writing and its Measurement

The arbitrary and often exclusive nature of collocations has been shown to be both important for academic writing and a challenge for those hoping to sharpen their skills in it. There has been a long-standing thesis in L2 literature that collocations pose significant challenges for second language users because of the difficulty of storing word combinations mentally in their entirety and being able to use them appropriately, given their often non-translation in a

learner's L1 (Nesselhauf, 2005). However, there is also evidence that collocations pose problems for first language users because they are expected to use them to signal their membership of a particular academic community (Wray, 2002; Durrant, 2019).

Attempts to tap into statistical collocations have used a word combination's frequency information or 'signature'. This frequency information has been used as a measure of exclusivity in the form of measures of association. Evert (2004) defines an association measure as a formula that computes an association score from the frequency information in a pair type's contingency table. These measures of association use the frequency of the first word in the combination, the frequency of the second word in the combination and the frequency of the combination itself and the total number of words in a large reference corpus to calculate one of two things: (1) the *statistical significance* of an association between the words *or* (2) the *degree* of association strength between the words in the combination. The scores are an indicator of how significant or how strong the association is. The more significant or strong, the more likely the combination is perceived to be a collocation.

Language learning literature has drawn on two representative measures from Evert's (2004) groups of *significance* and *degree* of association. In the former, the t-score has been used; and in the latter, the MI (Mutual Information) score has been used. Although, their formulae are discussed in Chapter Four (Section 4.4), these measures are able to illuminate different types of word combinations. Granger and Bestgen (2014) and Durrant (2019) make this clear by stating that the t-score provides information on how certain we can be that there is an association between the words i.e., the extent they appear by random chance. The t-score formula is known to illuminate word combinations that comprise of high frequency words. Examples of these include: '*other hand*', '*long time*', '*little bit*' (Granger & Bestgen, 2014). In contrast, as a measure of the strength of association, the MI formula emphasises

the exclusivity that the two words have with each other. Word combinations that have particularly high MI scores are known to comprise of low frequency words which have few alternative partners. Examples of high scoring MI combinations include: '*pop music*', '*juvenile delinquency*', '*vicious circle*', '*tectonic plate*' (Durrant & Schmitt, 2009; Granger & Bestgen, 2014).

The importance of this observation should not be overlooked for student writing. Several L2 focused studies have highlighted how these association measures flag up high or low frequency collocations that have a degree of exclusivity (mainly measured by MI); or how they have a strong association when tested with the t-score. Across EAP studies, Durrant and Schmitt (2009) found that L1 writers used more high scoring MI combinations (therefore more exclusive pairings) and that these combinations were also found in a narrow range of language domains (i.e., specialised academic disciplines) while non-native writers used more high scoring t-score combinations, which occurred in a wider range of language domains (e.g., found across genres and disciplinary areas). This finding has been corroborated to *some* extent[2] in other studies which have found more proficient writers use more exclusive pairings (e.g., Bestgen & Granger, 2014; Bestgen, 2017; Garner et al., 2018, 2019; Granger & Bestgen, 2014; Paquot, 2018, 2019). In brief, these studies have highlighted how the MI score increases with proficiency level (e.g., Garner et al., 2018, 2019; Paquot, 2018, 2019) however across individual contexts, this is *not* a straightforward linear increase across proficiency levels.

[2] Readers are encouraged to remember the nuanced differences in association measure unit of measurement (e.g., mean association measure score, threshold groupings of collocational status and proportions), language learning contexts across these studies and their focus on different types of writing which make consensus on these claims challenging.

Some studies (e.g., Granger & Bestgen, 2014) have found that higher proficiency levels use more high scoring MI combinations while lower proficiency levels use more high scoring t-score combinations. Bestgen and Granger (2014) also found a significant positive correlation between the MI and grade scores while they found a non-significant weak correlation between the mean t-score and grades. Bestgen (2017) is another study which again reinforces the idea that the MI has a positive correlation with grade scores while at the same time also reinforces that these patterns are not universal across assessment contexts. Across texts from the First Certificate in English (FCE) exam and the ICLE corpus of university students, he found a significant positive correlation between the mean MI and grades for both the FCE and ICLE texts however for the mean tscore, there was a significant positive correlation only for FCE texts.

Garner et al (2018, 2019) and Paquot (2018, 2019) are a series of CEFR focused studies that look at increases in collocation use across proficiency levels. With Korean learners, Garner et al (2018, 2019) found that there was a significant increase in the MI across A2-B2 CEFR levels when completing a university placement test while with French university linguistics assignments, Paquot (2018, 2019) found that increases in the MI were not linear across CEFR levels with increases often only present in higher CEFR levels.

Findings for L2 studies show a mixed trajectory that appear to support that higher proficiency levels use more higher MI combinations that are associated with academic writing and/or specific types of academic writing while lower proficiency levels use more higher t-score combinations that are associated with more generic types of writing however as a body of research, these patterns are not stable across proficiency levels or assessment contexts.

A smaller number of L1 or L1/L2 focused studies have also looked at collocations via association measures (e.g., Crossley et al., 2012; Durrant & Brenchley, in press; Durrant et al., 2019). In a study of L1 children's writing, Durrant and Brenchley (in press) looked at how children's collocation use develops across year groups. They found no significant variation in MI scores across age groups or genres across their corpus of literary and non-literary texts. In their study of first year university writing in different US universities, Crossley et al (2012) and Durrant et al (2019)'s studies again highlight the idiosyncratic nature of the relationship between collocations and grade scores. Crossley et al (2012) found a negative correlation between bigram and trigram frequencies from a reference corpus (the BNC) while in their study of multiple different projects at another US university, Durrant et al (2019) found a weak positive correlation between the MI and grade scores. In this sense then the use of both the MI and the t-score have taken on an important understanding for student writing. Through these measures, we can tap into the extent that students may use highly exclusive pairings that may be discipline or register specific or aspects of technical vocabulary (e.g., as in Durrant and Schmitt's (2009) example of 'tectonic plates'); and at the same time empirically examine their use of more frequent, everyday generic language use that is perhaps not particularly associated with academic written text.

Collectively, the existing studies that have looked at collocation-writing quality grade relationships present three observations that limit understandings of the collocation-grade relationship. First, that these findings appear to be contextually sensitive and do not hold consistently across the different contexts that have been studied. Second, that the existing studies mainly focus exclusively on the MI and t-score with only a few very recent studies (e.g., Durrant & Brenchley, in press; Durrant et al., 2019; Garner et al., 2018, 2019) using other available association measures. Third, the limited and often inappropriate choice of

statistical analyses and modelling techniques with many of these studies using correlation and multiple linear regression techniques.

1.3 Unpacking Contextual and Measurement Limitations

This section unpacks each of the three observations and in doing so explains the approach taken in the present study. Section 1.3.1 describes how existing studies highlight the context sensitivity of collocation development in writing and how this needs to be considered for the FYC contextual under study. Section 1.3.2 criticises the narrow range of association measures that have been used in existing studies and Section 1.3.3 finally highlights the limitations of using different statistical methods. Taken together these sub-sections provide the rationale for the study's focus and research design.

1.3.1 Context Sensitivity

Existing studies have shown some evidence that the MI increases with proficiency level and that the use of higher t-score combinations is associated with lower proficiency level. However, these increases are not uniform across assessment/learning contexts. Existing studies have therefore highlighted the different nature of collocation-grade relationships. Durrant and Brenchley (in press) comment that their findings with L1 children's writing that the MI does not increase linearly serves as a reminder that the concept of collocation development appears not to be uniform across contexts or learner groups.

This sensitivity is the first observation that the present study tackles. In the study of first-year university writing in the US, these FYC contexts are important settings worthy of further investigation because of their learner demographic including both L1 and L2 writers

who are assessed under the same programme outcomes as opposed to judgements of only language proficiency (as is the case of the purely L2 studies). This diverse setting is distinctly different from the research that has been carried out with purely L1 and L2 learners and therefore these past findings cannot be readily taken for granted as applying to mixed FYC contexts. This is the contextual gap that the present study aims to address.

1.3.2 The Existing Collocation Measure Set

The second observation that this study tackles relates to the number of association measures that have been used to measure the relationship between collocations and writing quality. Scholars have relied on a narrow set of measures that have been restricted to association measures used in the language learning/assessment literature with the t-score and MI featuring in all of these studies with sparse mention of alternatives or an awareness of how the hundreds of other association measures touted in the literature align with the MI or t-score and may be able to illuminate different collocation properties to those highlighted by the MI and t-score. The study will put forward the argument that the use of these measures needs to be justified and understood against the wider bank of association measures that researchers have access to and ultimately the measures need to be understood in terms of their formulas and their ability to illuminate different types of collocation properties. There is also a need to bring together the association measure literature which has been described as fragmented. This fragmented picture means measurement choice is often underexplored and/or under theorised because measures are spread out across different disciplines and scholars (Gries & Durrant, 2020).

The rationale for selecting measures like the MI has been openly questioned (e.g., Gablasova et al., 2017) and further still, the use of only these two measures seems to stand in stark contrast to the fact there are hundreds of association measures in the computational literature which can tap into similar and distinct properties of collocation (Gablasova et al., 2017; Pecina, 2005; Stevens et al., 1964; Wiechmann, 2008). Indeed, the narratives around measure selection in the existing collocation-grade studies often do not acknowledge where the MI and t-score fit into a landscape of possible association measures.

1.3.3 Choice of Statistical Analysis

A final observation that the study tackles focuses on how the relationship between features and writing quality has been measured. While studies in this research area have started to use a wider range of statistical methods (e.g., discriminant analysis (e.g., Crossley & McNamara, 2009), path analysis (e.g., Aryadoust, 2016), multidimensional analysis (e.g., Gardner et al., 2018)) and very recently mixed effects models (e.g., Durrant & Brenchley, in press; Paquot, 2018, 2019; Garner et al., 2018, 2019) to measure the relationships between features and writing quality, the research area continues to often inappropriately rely on correlation and/or multiple linear regression methods as found in our review of L1 and L2 corpus-based literature from 1945-2018 (Durrant, Brenchley & McCallum, 2021).

The field has tended to rely on these monofactorial methods to establish linguistic features that best explain or predict grade variation (Durrant et al., 2021). However, a key goal of the present study, that is developed across Chapter Four (Section 4.8) and Chapter Seven is to show how this type of approach is largely inappropriate for feature-writing quality

relationship research because it violates the statistical assumption of data independence. The present study will show how a mixed-effect modelling approach is a more valid method for studying hierarchical corpora whereby the corpus contains data dependency because more than one student contributes an essay to the corpus and each rater grades more than one essay. This means the data points making up the regression equation share a degree of dependency rather than being independent, as assumed by the use of monofactorial methods.

1.4 Contributions to Knowledge and Aims of the Study

These contextual and methodological limitations call into question what we have learned about the relationship between collocations (and linguistic features generally) and writing quality thus far. In tackling these limitations, the study contributes to the literature under four spheres of thought.

First, the investigation of collocation-writing quality relationships takes place in the under-studied context of First-Year Composition (FYC) writing programmes in the U.S. This choice of context contributes to the literature in two ways. First, the study directly compares L1 and L2 writers who complete the same tasks and are taught and assessed under the same programme objectives and evaluation criteria. This contrasts with much of what is already known about the relationship between collocation and writing quality as this knowledge has been largely acquired through the unidimensional study of L2 learners. A focus on FYC writing therefore offers a complementary contribution to those studies that have focused exclusively on L2 undergraduate writing (e.g., Granger & Bestgen, 2014; Bestgen & Granger, 2014), postgraduate L2 writing (e.g., Paquot, 2018, 2019), L2

placement test writing (e.g., Garner et al., 2018, 2019) or L2 exam-based writing (e.g., Bestgen, 2017).

Second and more broadly, the study contributes directly to a wider body of FYC literature which has started to call into question the lack of direct focus on language input on FYC programmes. These programmes are traditionally rhetoric-heavy instructional endeavours which several researchers have started to question (e.g., Aull, 2015a, 2015b, 2015c, 2017, 2019; Gere, 2016; Eckstein & Ferris, 2018; Perin & Lauterbach, 2018). Scholars in FYC contexts have started to point out that instruction and assessment should incorporate more explicit language instruction and guidance in their assessment criteria, more in line with EAP literature which embeds the teaching and assessing of language into its overall stance on composition and writing skills (Aull, 2015a). This study is therefore another voice in this developing FYC narrative that unpacks potential areas of linguistic focus for FYC instructors. The importance of such a study is clear in that establishing or not establishing a link between collocations and writing quality grades will help add objective empirical weight to the FYC scholarly voices which have so far supported a shift from process instruction towards an approach which incorporates language use.

The third and fourth contributions relate directly to the methodological approach taken. The third contribution lies in the use of a cluster analysis to unpack the relationships that exist between the many association measures that the literature has referred to. The cluster analysis establishes which measures are similar/distinct and able to flag up different types of collocation properties. This cluster analysis therefore brings together and evaluates many of the measures of association that appear fragmented across the literature.

The fourth and final contribution is made in the way that the study investigates the relationship between collocations and writing quality, not by relying on traditional multiple linear regression modelling, but by using mixed-effects modelling to take account of the nature of the study corpus and the assessment context. The use of mixed effects modelling takes into account both fixed and random effects and therefore measures relationships more robustly by accounting for random corpus and contextual effects which account for variation in the relationships.

In making these contributions to knowledge the study aims to gain an understanding of the relationship between collocations and writing quality through the following:

(1) Performing a cluster analysis on a range of association measures so as to uncover and explore the number of these measures which seem to be highlighting different, varied properties of collocation,

(2) Examining how the resulting association measures in the cluster analysis have a relationship with writing quality grades,

(3) Increasing the validity of measuring collocation-grade relationships by factoring in several relevant learner and contextual effect variables that are thought to influence such measurements,

1.5 Research Questions

The study is guided by the following broad research questions:

RQ 1: To what extent do the vast array of association measures help tap into different collocation properties? and how can we use this information to select principled measures for collocation analysis in a language learning context?

RQ 2: To what extent do measures of collocation have a relationship with writing quality?

RQ 2.1: When a set of these association measures are selected, to what extent do they have a relationship with writing quality?

RQ 2.2: To what extent do the fixed effects of task and language status also have a relationship with writing quality?

RQ 2.3: To what extent do these relationships vary when the modelling process considers the random effects of individual rater and/or individual student?

1.6 Organisation of the Study

The remaining seven chapters of this study proceed as follows.

Chapter Two sets out the FYC context that informs the study and outlines the programme's objectives and its approach to teaching and assessment. The chapter also describes how the programme uses an online Learning Management System (LMS) to review and give feedback on student essays as well as provide grade breakdowns across its two modules.

Chapter Three reviews the study of the relationship between measures of vocabulary and writing quality. Chapter Four connects the approaches taken to the study of vocabulary in Chapter Three with those approaches taken more recently in the study of phraseology and more specifically the study of collocations. In addition to reviewing these studies, the chapter illuminates methodological concerns with the use of simple monofactorial research methods as well as details how contextual and sampling variables have influenced the measurement of such relationships.

Chapter Five gives an overview of the study methodology including the research design and ethical considerations, how the FYC corpus was constructed, and how texts were annotated prior to analysis.

Chapter Six presents the initial cluster analysis study by describing the measures in the analysis, how the cluster analysis was carried out and provides an interpretation of the results that go on to inform the subsequent mixed-effects model study.

Chapter Seven describes the interactions between the contextual and learner variables in the relationship between collocations and writing grades. The chapter presents findings that explain the interactions between random effect variables and how these findings have implications for assessment on the FYC programme.

Chapter Eight concludes the study by summarising the main contributions to knowledge and how these contributions have important implications for the teaching and assessment of FYC writing. The chapter ends by acknowledging the limitations of the study and how future work can build on the results presented.

Chapter Two: Context

2.1 Introduction

The chapter begins by outlining the general philosophy of FYC programmes and their key goals and then moves on to consider the specific objectives and approach taken at the University of South Florida (USF). The chapter covers individual module information as well as provides an overview of how the modules are connected. This context chapter is an essential element of the study, because, as will become apparent in the Chapter Five's methodology overview, the programme set up contributes to the procedures undertaken in the research design, data collection and analyses.

2.2 First Year Composition Programmes in the U.S

FYC programmes are a key component of the undergraduate curriculum at US universities. The Council of Writing Programme Administrators (CWPA) set out general guidelines on FYC programmes that institutions follow. The CWPA Outcomes Statement (2014) clarifies the broad set up and goals of these programmes. The programmes typically operate as sequences of modules that develop students' knowledge of rhetoric by exposing them to a range of text types, which they are also then expected to produce throughout their enrolment on the programmes. A key cornerstone of this exposure relies on developing an awareness of genre conventions, formatting and citation practices, and developing the ability to use literature to create arguments or set out key stances expressed within their chosen topic.

The CWPA Outcomes Statement (2014) elaborates on the skills students need to exercise and acquire by drawing attention to students' abilities to change tone, voice, and formality. At the same time students are expected to structure their written texts in a way that both meets the different aims their different text types have and develops their ability to write for specific audiences. There is also a focus on reading and critical thinking skills, which aim to allow students to selectively include key views from their reading when writing their own texts (CWPA, 2014; Eckstein & Ferris, 2018).

In setting out the key skillsets that students develop, the CWPA Outcomes Statement (2014) and the CWPA (2014) more widely regulate FYC programmes by relying on guidelines that standardise aspects of the programmes across US universities. However, Eckstein and Ferris (2018) point out that while the interrelated Conference on College Composition and Communication (CCCC) statement (2014) promotes addressing students' linguistic needs, the loose nature of the guidelines means there is great potential for language input and explicit instruction to be overlooked on FYC programmes in favour of focusing more on writing processes. Eckstein and Ferris (2018) rightly acknowledge how this negligence may have the greatest impact on L2 learners as they struggle with peer review, critical thinking and text ownership skills. However, while this is indeed a valid observation, there is also scope for L1 students, who all enrol from different high school backgrounds, to also need support in these areas (Matsuda, 2006). It is important to remember that often, L1 students do not have enough academic English experience upon entering university to cope with the demands of university level study or assessment (Bychkovska & Lee, 2017). Bychkovska and Lee (2017) state that academic writing is not a native language for anyone; however, as will be detailed in this study's literature review (See Chapter Four Section 4.5), many of the studies that have looked at language use across

grade levels have done so from a vantage point that considers L1 writing as a benchmark for L2 writers. In such a view language use from L2 writers is thought of as deficient, when compared with their L1 counterparts.

These struggles are illuminated further as some FYC programmes give little explicit language instruction despite evidence that coursework grades are reduced for language related problems (Anson, 2000; Matsuda, 2012; Matsuda et al., 2013, cited in Eckstein & Ferris, 2018).

2.3 The University of South Florida

The University of South Florida is a large, public university with campuses across 3 locations in the state of Florida: Tampa, St. Petersburg and Sarasota-Manatee. Students are required to pay tuition fees irrespective of undergraduate or graduate status however, in recent years, most enrolled students have received some form of financial aid (Points of Pride at USF, 2018). The university is home to a large population of students (circa 50,000) of varying demographic backgrounds with as much as 41% thought to identify as African American, Black, Asian American, Hispanic, Native American or multiracial (Points of Pride at USF, 2018). The university also caters for a range of academic disciplines including but not limited to business, engineering, arts and social sciences and interdisciplinary sciences as well as having a strong athletics culture (Points of Pride at USF, 2018). Undergraduate first year students who join the university are accepted on merit with students having an average Grade Point Average (GPA) of 4.12 and an average SAT score of 1280. These two averages exceed the benchmarks set by the State of Florida Board of Governors (Points of Pride at USF, 2018).

2.4 USF's FYC Objectives, Structure and Student Population

This section of the chapter outlines the programme objectives, structure and student demographic.

2.4.1 Connections between General Education and FYC

The state of Florida has two main general education requirements for undergraduate degree programmes: State Mandated Core and State Required Communication and Computation (formerly known as the 'Gordon Rule'). Students who enter a Florida College State System or Florida State University System have, from 2015/2016, been required to complete 36 hours of general education coursework from a list of courses in subjects including communication, mathematics, social sciences, humanities and natural sciences[3]. Students must complete one course from each subject area. These guidelines are set out in the Florida Board of Governors Regulation (Regulation: 8.005). These regulations ensure that students are equipped with academic-level literacy and numeracy skills that will allow them to cope with the general demands of the content-focused curriculum they meet later in their degree programmes.

[3] At the time of writing, these were the regulations applicable to the writers who produced their FYC texts in the academic year 2016/2017. These regulations also included mathematics requirements, satisfying the State Required Computation component in that students were required to take either 3 or 6 credit hours of Mathematics and/or 0-3 credits of quantitative reasoning under the specified 36 hours .

Courses approved to meet the USF State Required Communication enable students to demonstrate college-level writing skills through multiple assignments. Students will engage in writing as a “process”, which means employing strategies such as pre-writing, co-authoring, document design, peer feedback, revising and editing. Overall, students will learn how to develop ideas and texts that follow academic/ disciplinary conventions for different contexts, audiences and purposes. These learning outcomes are assessed through written assignments that include essays, creative writing, journals, written examinations, portfolios, case studies, letters and proposals. This requirement is met at USF through its FYC programme, which provides students with a programme that meets the State’s requirements. Students need to achieve a C- grade or higher to satisfy the requirements of the FYC programme and more broadly the State’s general education requirements.

2.4.2 The FYC Programme at USF

The FYC programme at USF is among the largest in the US with student enrolment per academic semester reaching around 4,500 students (Moxley, 2013). In line with other FYC programmes, all first-year undergraduate students are required to take the FYC programme unless they have exemptions through taking similar courses including achieving high SAT or English Advanced Placement (AP) exam scores (Alaina Tackitt, personal communication, 2018; Durrant, Moxley & McCallum, 2019). USF allows students to receive up to 45 semester hours of credit towards the baccalaureate degree upon successful completion of a range of examinations. Credits earned from international institutions or study abroad programmes are also evaluated for transfer. In the case of the FYC programme, students can be granted an exemption if they have taken the English AP examination. Guidelines

specify that if a student scores a 4 or 5 on English Language and English Literature examinations they can receive exemptions from the two English composition ('ENC': English Composition) programme modules titled 'ENC 1101' and 'ENC 1102'. However, a score of 3 in these exams will only result in an exemption from ENC 1101 (Credit-by-Exam Equivalents, 2018).

2.4.2.1 The FYC programme and module learning outcomes

Students are required to meet a wide range of programme objectives, which include:

- Learning and applying effective strategies that advance critical reading, drafting, reviewing, collaborating, giving of helpful peer feedback, revising, rewriting, rereading and editing skills,
- Summarising research sources through effective annotation, note-taking, quotation, citation and paraphrase,
- Composing several academic genres and adhere to their academic conventions including structure, citation and linguistic features,
- Identifying and developing organisational strategies that contribute to the effective delivery of information and argument,
- Demonstrating responsiveness within an established disciplinary context to new information, experiences and ideas through a process of re-evaluating ideas and/or approaches,
- Analysing rhetorical effectiveness from a variety of print and non-print sources,
- Evaluating relevant sources according to their contexts, rhetorical situation, usefulness and credibility for specific research tasks,

- Synthesising disparate or conflicting thoughts when evaluating questions/problems to form cohesive and collaborative solutions,

These learning outcomes are applicable to the two programme modules: ENC 1101 and ENC 1102.

The programme's two modules require students to write three assessed projects per module meaning students write a total of six projects on the programme[4]. Table 1 details these six projects:

[4] The remaining percentage of the final grades comprises of other FYC programme activities and students also receive a percentage for participation/attendance related activities.

Table 1: Overview of FYC Projects

Module	Project	Coursework assessment	Student Tasks	Project grade
ENC 1101	**1**	Annotated bibliography	(a) Recognise an argument by citing, outlining and summarising a public writing source (performing preliminary research) (b) Practice peer review skills by evaluating a published article (evaluating research) (c) Compile research and narrowing the research area (creating a process log) (d) Write annotations, making connections, and identifying stakeholder interests (formalising research)	20% of final grade
	2	Analysing a stakeholder's platform	(a) Analyse a stakeholder's website as an early draft (b) Interpret, explain and evaluate the stakeholder's platform by writing an intermediate draft of analysis (c) Practice peer review skills by responding to peers' intermediate drafts (d) Map a revision plan which makes analysis stronger (e) Finalise analysis of the website	25% of final grade
	3	Synthesizing multiple perspectives in a literature review	(a) Recognise synthesis of multiple sources by creating an outline as an early draft (b) Advance the synthesis by writing the intermediate draft of the literature (c) Practice peer review skills by responding to peers' intermediate drafts (d) Map a revision plan to improve the synthesis	30% final grade

ENC 1102	**1**	Rogerian argument: Finding Common Ground	(a) Focus research by responding to guiding questions in an early draft (b) Formalise research by writing an intermediate draft (c) Practice peer review skills by responding to peers' intermediate draft (d) Map a revision plan to improve Rogerian argument	20% of final grade
	2	Analysing visual rhetoric	(a) Focus research by responding to guiding questions as an early draft (b) Formalise research by writing an intermediate draft (c) Practice peer review skills by responding to peers' intermediate drafts (d) Map a revision plan to improve the visual analysis (e) Finalise the visual analysis	25% of final grade
	3	Composing a multimodal argument	(a) Prepare a preliminary response that comprises of composing a multimodal argument (b) Educate, enhance and empower a non-engaged stakeholder by composing an intermediate draft of multimodal argument (c) Practice peer review skills by responding to peers' intermediate drafts (d) Map a revision plan to improve the multimodal argument (e) Finalise multimodal argument	30% of final grade

Across these two modules, instructors work through a standardised curriculum that has been developed collaboratively between the FYC administration, faculty and GTAs (Graduate Teaching Assistants) (Dixon & Moxley, 2013). Over a 15-week semester, ENC 1101 aims to solidify writing practices by introducing and practicing the key skills of paraphrasing, citing sources, drafting and editing and peer collaboration and feedback. In ENC 1101, these aims are guided by using an online e-book : '*Rhetoric Matters: Foundations of Rhetoric and Composition*' (Hoffman & Wiggs, 2016a), which is published in-house and is available to students via the programme's online My Reviewers[5] (My R) platform.

Students complete an annotated bibliography, a thesis-driven essay and a remediation essay as their 3 projects, which are weighted as contributing to 20%, 25% and 30% respectively of their overall module grade. The second and third projects are linked because the remediation essay is an 800-1000-word essay remediating the second essay project into a multimodal format (Dixon & Moxley, 2013). The assignment is termed 'remediation' because it expects students to transform their writing from one genre into another with the new genre (the multimodal format in this FYC case) expected to have rhetorical benefits for readers of the work. The remaining 25% of the module grade is allocated as 10% for attendance and 15% for classwork and homework. This classwork and homework takes the form of quizzes on aspects of writing including citation practices and avoiding plagiarism. Students may receive a fail grade if their attendance falls below 67% for the module (USF ENC 1101 syllabus).

[5] The MyReviewers platform can be accessed at: http://myreviewers.usf.edu Last accessed: 05.05.2020.

ENC 1102 develops students' writing abilities by focusing on how writers negotiate various contexts via argument and reasoning (USF ENC 1102 Syllabus, 2018). ENC 1102 looks more closely at developing agency in a text with each of its three projects focusing on different but equally necessary components. Similar to ENC 1101, the core components needed to produce the module's projects is taught by using another in house created e-book: '*Rhetoric Really Matters': Explorations and argumentation'* (Hoffman & Wiggs, 2016b). Project 1 requires students to develop arguments that negotiate differences in the views of stakeholders. Project 1 therefore is intended to build on the last project of ENC 1101 (which had already simply described the key arguments around the topic) because it requires students to critically analyse these key arguments and reach a compromise. Project 2 requires students to analyse how visuals are used to portray arguments and project 3 brings together these text and image-based contributions in the form of creating a multimodal argument. Within ENC 1101 and ENC 1102 students are expected to receive multiple rounds of instructor and peer feedback via My Reviewers (My Reviewers, 2018).

It is clear that across the two modules, the level of difficulty is intended to increase with students moving from summary-synthesis-argument in their research-informed writing, to using these sources in increasingly complex ways across the projects in terms of generating arguments and commenting on different argumentative strands found within the literature (Moxley & Eubanks, 2015).

The next section of the context chapter examines how the construct of writing proficiency has been presented and measured across different types of academic student writing, how the same construct has been presented and measured in the USF FYC programme and importantly how this approach to

measuring writing proficiency is carried out on the programme with input from students and instructors.

2.5 Understanding Writing Evaluation on the FYC Programme

The majority of FYC instructors are GTAs who are Doctoral candidates. The pool of instructors also includes adjuncts and visiting instructors who teach one to four classes or sections each . The GTAs attend a three-day induction if they have taught on the programme before; or a five-day induction if they are new to teaching on the programme. The induction includes workshops on grading practices and becoming familiar with the grading rubric. A mentor is assigned to each new instructor with the mentor responsible for reviewing the new instructors' comments on students' work and their grade allocations (Tackitt et al., 2016).

Instructors are also supported via a materials bank including resources such as marked up sample essays, extensive banks of common comments, videos and articles about the grading rubric criteria which are available on My Reviewers, and the university's Canvas platform which stores programme documents and grade administration (Alaina Tackitt, personal communication, 2018; Tackitt et al., 2016). Taking the grading and generic course information together, the general process pedagogy approach that is used throughout the FYC programme can be said to be cyclical in nature in terms of the modules, whereby alternate weeks in the module syllabi are devoted to holding writing conferences with students to discuss ideas and receive guidance, peer review completion and producing and implementing revision plans to produce final drafts that meet the project's task requirements[6].

[6] This scaffolded and collaborative approach is in part facilitated by having class sizes of 19 students of both L1 and L2 speakers who are expected to play an active role in peer reviewing each other's essays.

Each project has its own rubric, which is divided into analysis, evidence, organisation, focus and style components. All grades are awarded for these components on a 0-8 scale (shown in Appendix A) with an overall holistic grade also awarded, which is scored on a 15-point scale expressed by the letters A-F (shown in Table 2) (Durrant et al., 2019; Tackitt, et al., 2016). We can see in Table 2 that the grades A-F correspond to the Grade Point Average (GPA) with an A+ equalling a GPA of 4.0 (the maximum GPA awarded in the US system).

Table 2: Holistic Grades Awarded for ENC 1101 and ENC 1102

Grade Types	**Grade breakdown for ENC 1101 and ENC 1102**		
A grade	A+ (97-100) 4.00	A (94-96.9) 4.00	A- (90-93.9) 3.67
B grade	B+ (87-89.9) 3.33	B (84-86.9) 3.00	B- (80-83.9) 2.67
C grade	C+ (77-79.9) 2.33	C (74-76.9) 2.00	C- (70-73.9) 1.67
D grade	D+ (67-69.9) 1.33	D (64-66.9) 1.00	D- (60-63.9) 0.67
F grade	F (59.99 or below) 0.00		

ENC 1101 uses the same rubric for all three of its projects with a focus on 5 core criteria: focus, evidence, organisation, style and format. The focus, organisation and style categories are divided into basic and critical thinking sub-components. The format category is considered basic while the evidence category is considered critical thinking. This means instructors provide a grade for a total of 8 subcategories: Focus (the two sub scores R1 and R2), Evidence (R3), Organisation (R4 & R5), Style (R6 & R7), and Format (R8). In ensuring that instructors and students understand the rubric and the programme grading procedures, both parties have access to videos that explain the rubric terms and

sample marked up papers that explain why such grades were awarded by instructors.

ENC 1102 also uses the same 0-8 component scale and A-F holistic scale; however, the textual descriptions under each component pay attention to slightly different aspects of writing to reflect the different goals that each module has. The ENC 1102 rubric is presented in Appendix A. For example, more focus is given in ENC 1102 Project 1's rubric to Rogerian style of argument (under the analysis criterion) while in ENC 1101's Project 3, the focus is simply on communicating a thesis statement.

2.5.1 Using the My Reviewers System for Grading and Feedback

The FYC programme uses its own suite of writing tools under the name of 'My Reviewers'. My Reviewers was first developed in 2009 as a response to criticisms that generic commercial writing tools did not meet the specific needs of USF's first year student population[7] (Branham, Moxley & Ross, 2015). The My Reviewers interface is shown in Figure 1.

[7] The My Reviewers platform has since been adopted in several universities across the US including the University of Pennsylvania and North Carolina State University.

Figure 1:My Reviewers Interface

Moxley and Eubanks (2015) and Moxley (2012) underline how the philosophy behind My Reviewers views text creation as a social practice through which students can provide valuable feedback to one another based on their backgrounds as readers and critical thinkers. As a platform, My Reviewers has the primary aim of improving students' writing, critical thinking skills and collaborative experiences. In facilitating these aims, My Reviewers explicitly aims to provide feedback on student writing and improve peer feedback practices. The My Reviewers platform achieves this aim by using of a multimedia library of over 200 instructor-created community comments that offer general advice, grammar and mechanics tips, and additional resources that address common writing concerns (Branham, et al., 2015; Moxley & Eubanks, 2015). Each comment links to an article, video and/or exercises that inform students about the area of writing practice with which the comment is associated. Upon clicking on these individual

comments, students are taken to a help interface that contains examples that are of a similar nature to their own writing. They are prompted to rectify the particular writing concern that has been flagged up through a series of questions.

Table 3 includes a selection of these comments and guiding questions to support students in their self-revisions. In addition to the examples in Table 3, other comments focus on aspects of sentence variety and/or sentence structure as well as word order problems. In the case of word order, this comment explores how word order affects meaning through placement of 'only':

(1) 'He ***only*** said he loved her' and

(2) 'He said ***only*** he loved her'

The first example implies he just said it perhaps he does not mean it and the second example implies there is an assumption that no one else loves her.

Table 3: Community Comment Examples

Aspect of writing	Community Comment example	Guiding questions/explanation that encourages self-revision
Language use	Diction/Word Choice Problem[8]	What is a diction problem? A diction problem happens when you use a word in the wrong context or use a word that does not mean that you intended it to mean in that situation. Diction or word choice always depends on context.
Argument	Weak argumentation	There are many reasons why a writer's argument may be weak. Here are just a few possibilities: • The author's argument is undermined by one or more logical fallacies • The author does not adequately consider the rhetorical situation at hand.

[8] In each case here, students are also directed to a possible Youtube video and/or online resource. For example, in the case of word choice/diction, students are given a link to a phrasal verb dictionary.

Organisation	Logical organisation[9]	Why is it important to organize a paper logically? For the sake of clarity and cohesiveness, a logical plan should inform the paper's organization from beginning to end at both the global and local levels. The target audience is more likely to become engaged, and maintain their engagement, when the conversation is clearly organized and purposefully presented.

[9] Students are presented with suggestions of planning and organisation techniques including mind mapping and outlining at paragraph level.

These comments also help guide students in their peer review reading by providing models of the type of comments they may make on their peers' drafts.

Moxley and Eubanks, (2015) and Moxley (2012) note how the PDF mark-up tools allow instructors to give students extra support by using endnotes and hyperlinks to articles. Further to this, My Reviewers can track students' progress on a specific learning outcome as well as assign, conduct or grade peer reviews. The My Reviewers Student Manual (2018) includes tutorial videos that guide students through several activities in which they are required to participate. These include setting up their My Reviewers account, uploading their assignments, revision plans, assessing community comments, recording audio comments, completing peer review reports and uploading videos.

2.6 Instruction on Language Use in the FYC Programme

Matsuda (2006) points out that FYC programmes have a tendency to overlook the specific needs of L2 writers and in doing so instruction does not often target or address the needs they have. However, as Aull (2015a) and Connors (1997) indicate the original need for FYC programmes arose because native speakers also struggle (and continue to do so) with essay writing and college-level tasks as their writing skills are below the standard expected of college students.

Aull (2015a) recognises that most FYC programmes operate under process pedagogies that champion planning, drafting and editing while tasks require students (L1 and L2 populations) to produce audience and genre specific texts with a neglect for how language use facilitates meeting these aims. Matsuda (2006, p. 640) encourages instructors and WPAs to remember that: "Language issues are also inextricably tied to the goals of college composition, which is to help students become better writers". Matsuda (2006) explains how most

instructors' definitions of "better" involves the ability to produce language that is unmarked and appropriate for the writing context/ situation. Zawacki and Habib (2014) also note how students struggle to achieve this unmarked language use because they fail to understand how style and language forms combine to create texts that meet the readership needs of a specific community or discipline of practice.

Matsuda (2006) and Aull (2015a) also acknowledge that language is explicitly included in first year assessment criteria as well as pre-university level. Both L1 and L2 populations are expected to show skilful use of language and vocabulary and language use that is free from grammatical, mechanical and usage errors in national and international levels of assessment (Aull, 2015a, p.8). This is evident in US contexts, including national SATs and international assessments such as TOEFL and IELTS which serve as proficiency level entry checks into undergraduate and graduate levels of education (Aull, 2015a). Gere (2016) further explains in the review of Aull's (2015a) work how Aull (2015a) captures the importance of using appropriate language-level patterns and how these patterns do not operate in isolation from macro-level writing processes such as argumentation goals and use of evidence but instead help facilitate these processes.

On the surface, ENC 1101 and ENC 1102 instruction appears to be primarily focused on the processes of writing; however, as the 'Community Comments' in Table 5 show, there is explicit focus on local language concerns including aspects that have direct relevance to the phenomena of collocation. These include word choice and diction. There is further clear reference to language use in the programme's self-published 'My Reviewers guide to style and grammar' (Downey, Loyer, & Walkup, 2016). The guide focuses on several

aspects of language that is directly relevant to students writing up draft and final project essays. These aspects include active vs. passive sentences, combined and awkward sentences, the art of being concise, diction and word forms.

In terms of how these nuances of language are reflected in assessment, there is evidence that they play a role in allocating grades across the module projects and their assessment rubrics. This reference can be seen under the evidence component, where texts that fail to achieve pass grades rarely distinguish between writer's ideas and sources ideas and quotes, paraphrases and summaries are not clearly and consistently introduced, integrated and analysed to support arguments (See Appendix A, p.245). Arguably, the clearest reference to language appears under the style component where in weak texts "language significantly interferes with communication of ideas with frequent grammar and/or punctuation errors, inconsistent points of view, significant problems with syntax, diction and word choice" (See Appendix A, p. 246). These descriptions, which guide project raters, help indicate that while the programme centres on process pedagogies, a focus on language-level features is not wholly absent as writers are expected to conform to appropriate academic word choice in keeping with the aims of producing academic texts for a specific reader. Writing Programme Administrators (WPAs) within the FYC programme further highlight the role of language in that some instructors' material draws attention to specific language choices in published articles or sample essays. They also note that Instructors have a degree of flexibility in what material is used and many expose students to sample previous projects whereby students may acquire and produce aspects of language that are similar to that in the texts they are exposed to (Alaina Tackitt, personal communication, 2018).

In line with this acknowledgement, from the CWPA Outcome Statement (2014), we see how achieving certain rhetorical goals/functions is a key learning outcome for FYC programmes. While there is an undertone of how language plays a role in students meeting FYC programme outcomes, it is perhaps clearest when we examine the development of '*Rhetorical Knowledge*' and '*Knowledge of Conventions*'. In developing rhetorical knowledge, students are expected to develop facility in responding to a variety of situations and contexts calling for purposeful shifts in voice, tone, level of formality, design, medium and/or structure. In developing this knowledge, instructors are expected to guide students towards learning the expectations of readers in their fields, the main features of the genres in their fields and the main purposes of composing in their fields.

Reference to language becomes more prominent when the Outcomes set out their '*Knowledge of Conventions*' guidance. The Outcomes Statement defines conventions as the formal rules and informal guidelines that define genres, and in doing so, shape readers' and writers' perceptions of correctness or appropriateness.

Under this knowledge, the CWPA Outcome Statement expects students to:

- Develop knowledge of linguistic structures, including grammar, punctuation, and spelling, through practice in composing and revising,
- Understand why genre conventions for structure, paragraphing, tone, and mechanics vary,
- Gain experience negotiating variations in genre conventions,
- Learn common formats and/or design features for different kinds of texts,

- Explore the concepts of intellectual property (such as fair use and copyright) that motivate documentation conventions,

In setting out these student goals, instructors are expected to help raise awareness of:

- The reasons behind conventions of usage, specialized vocabulary, format, and citation systems in their fields or disciplines,
- Strategies for controlling conventions in their fields or disciplines,
- Factors that influence the ways work is designed, documented, and disseminated in their fields,
- Ways to make informed decisions about intellectual property issues connected to common genres and modalities in their fields,

Although these statements help clarify the role of instructors and students in this development of knowledge, there is little explicit guidance to help students develop and acquire the language resources to meet/perform these rhetorical goals. A point worth addressing here is that this lack of language focus impacts on all students enrolled on FYC programmes. Aull (2015a) and Hyland and Guinda - Sancho (2012) explain that irrespective of language background, students are not accustomed to the staged project work and writing for a field specific audience that FYC programmes require. Jeffery and Wilcox (2013) elaborate on this by highlighting that with US high school students, their NEAP exam requires students to write opinion-based arguments about large-scale phenomena whereby their opinion is supported by personal evidence. Similarly, international L2 students often encounter IELTS and are asked to consider topics like whether or not tradition and technology are compatible (Moore & Morton, 2005). Aull (2015a) contrasts these topic centred tasks with the source-based

arguments that students are expected to produce at university. Scholars argue that a focus on language in FYC writing can transfer beyond first-year writing courses (Granville & Dison, 2005; Tarratt et al., 2009; Negretti, 2012).

Given the recognised shortcomings of FYC curricula and instruction, researchers have started to broaden the scope of their studies by drawing on corpus-based analysis techniques that have frequented other learner writing contexts such as the EAP focus that is popular in Euro-centric Higher Education (e.g., see the overview in Aull, 2015a). It is this focus that has shaped much of the previous language-oriented studies that have started to be conducted in FYC programmes.

Gere (2016) highlights the originality and contribution that this corpus-based work has for FYC programmes when reviewing the early work of Aull (2015a). Aull (2015a, 2015b) uses a corpus of 19,463 FYC texts to ascertain how FYC texts differ from published academic writing. Aull (2015c) looks at the extent to which the U.S Common Core Standards can be used to clarify the connection between the arguments students (are expected to) make and the language/discourse resources they have in their repertoires to construct these arguments. Aull (2017, 2019) builds on her key argument that language use must be understood and incorporated into FYC rubrics and programme learning outcomes. Aull (2017) examines the variation in language use across FYC genres with a corpus of USF texts by looking at differences in the keywords that each genre's texts contain, whereas Aull's (2019) latest work compares undergraduate and graduate writing in the MICUSP corpus in terms of stance markers and notes differences in stance marker use across levels of study and genres (argumentative and explanatory writing).

While Aull's work has relied on traditional corpus-based tools and methods in the form of frequency counts and keyword analysis, we can see a shift towards using automated tool analysers in the work of Eckstein and Ferris (2018)[10] who also examine FYC writing. There is also a research shift with the work departing from the descriptive information that Aull provides to more inferential statistics that are geared to test for significant differences between language use and grades awarded. Eckstein and Ferris (2018) compare L1 and L2 FYC writers across the syntactic and lexical features contained within the Syntactic Complexity Analyser (SCA) and Lexical Complexity (LCA) tools from Lu (2011) and Lu and Ai (2015). Although, the focus of the present study is on collocation, these tools are best understood as providing frequency-derived counts of features including types, tokens, verb variation and noun phrase length. On comparing the L1 and L2 groups, among the key findings are that L1 writers have greater lexical variation.

Although these emerging studies undoubtedly help present a clear picture of student language use on FYC programmes, there are a number of avenues that FYC researchers should consider in the future. It is clear that these researchers have access to large amounts of data and future research foci should be on taking advantage of this data so that they can make more connections to exploring actual language use and teaching and assessment practices on FYC programmes. In the case of teaching and assessment practices, there are still numerous unanswered questions as to how individual writers use language and

[10] Perin and Lauterbach (2018) is another such study that follows this automated route. However, since the writers are studying at a community college, the study is not discussed further here.

importantly how individual raters view this language use when judging project tasks.

With this in mind, the present study aims to further highlight the role language plays in first year writing by examining the relationship collocation has with the construct of writing quality. In doing so, a number of contextual and learner variables that relate to the writing task and the language status of the writer are also given consideration in tapping into this relationship so as to appreciate the natural learning and assessment context at USF.

2.7 Summary

The next chapters proceed as follows. Chapter Three outlines the traditional study of the relationship between vocabulary and writing quality, Chapter Four uses this foundation to explain how research has moved on to study how phraseological complexity has a relationship to writing quality and how this relationship can be better measured by factoring in differences across tasks, language status and individual rater variables.

Chapter Three: Relationships between Vocabulary and Writing Quality

3.1 Introduction

The recent study of the relationship between collocation and writing quality has its roots in a wider body of work, which has traditionally involved the study of lexis at individual word level. The review that follows begins by outlining why a specific focus on lexis has been undertaken, followed by an examination of the connection between the theoretical constructs of vocabulary knowledge and the measures that have been used to understand how this knowledge comes to light in written text production. The purpose of this chapter is to contextualise the study of collocation within its informing body of single word research and to show how the approaches taken in single word work also relate to the approaches taken to collocation.

3.2 The Importance of Vocabulary in Determining Writing Quality

The importance of vocabulary to first and second language learning has been attested since Roman times, when the Romans were presented with lists of words organised by topic to aid their learning of Greek (Schmitt, 2000). In contemporary society, the importance of vocabulary is well-recognised by practitioners and students alike with instructors perceiving lexical errors to be more serious than grammatical errors in impeding message communication (Ellis, cited in Schmitt, 2000; Santos, 1988). Learners are also quick to equate their poor written performance with poor vocabulary production, which impedes their ability to formulate and communicate their ideas effectively (Llach, 2007). In further acknowledgement, Carter and McCarthy (1988) set out how learners face more challenges learning and using vocabulary than grammar because vocabulary is in a constant state of movement with words being coined, dropped from usage

and reinvented by language users. They note that the same changes do not apply so rapidly to syntactic structures because their much more rigid rule-based system means these structures remain familiar and constant.

Leki and Carson's (1994) assertion that academic success is related to vocabulary size also supports this earlier work. While Llach (2007) further extends the importance of vocabulary directly to student desires by highlighting that having a large bank of vocabulary is seen as a primary goal for students and they believe that, if lacking, their development and success stalls. Crossley and McNamara (2011), Olinghouse and Wilson (2013), Llach (2007) and Engber (1995) and work from the English Profile (e.g., Hawkins & Filipovic, 2012; North, 2014; Harrison & Barker, 2015) project have all pointed out the value of vocabulary and how the vocabulary choices a writer makes have an impact on the numerical grade score a text receives when assessed as part of coursework or a placement test. Text evaluators remark that second language learners have noticeable problems with vocabulary as they can tend to choose 'basic' choices whereas there is evidence that a competent native speaker tends to opt for a more precise, lower frequency choice (Schmitt, 2000).

The next sub-sections of this chapter set out in some detail how vocabulary knowledge has been unpacked and how its dimensions have been operationalised in the literature.

3.3 Vocabulary Knowledge across Learning Contexts

The construct of vocabulary proficiency is underpinned by the notion of breadth and depth of vocabulary knowledge (Anderson & Freebody, 1981). Breadth relates to the number of words a person knows and depth relates to what they know about each individual word with a person expected to have a sufficiently deep understanding of individual words (Anderson & Freebody, 1981, p.22). Under these definitions, it may seem rather obvious and simple to envisage that breadth relates simply to the number of words a person knows, however the conceptualisation of vocabulary depth involves a wider range of components. Depth can include knowledge of *phonology*, knowledge of the *morphemic* features of a word, knowledge of *syntax* and *semantics* (the meanings the word has and how it relates to other words), the word's *collocational* features or properties and its *register* features (knowing the contexts where the word is typically used) (e.g., See the overview in Durrant et al., 2021; Read, 2000; Nation, 2013). It is clear then that breadth and depth involve a complex array of knowledge types and in turn these types of knowledge have been measured in a number of ways when it comes to objectively quantifying relationships between vocabulary and writing quality.

In terms of looking at writers' texts, aspects of breadth and depth are judged by way of looking at the vocabulary that writers have produced in their writing. Although Durrant et al (2021) comment that knowing about either breadth or depth on their own is meaningless, studies that have measured the relationship between lexis and writing quality have often only studied one half of this knowledge base (i.e. either breadth or depth). Figure 2 shows the ultimately blurry connections between the theoretical constructs of breadth and depth, their match up to the operational constructs of Read's (2000) lexical richness:

diversity and sophistication. Figure 2 also shows an overview of quantitative measures that have typically been classified as either measures of diversity or sophistication. Figure 2 highlights that the theoretical constructs of knowledge have been pinned down to Read's (2000) notion of lexical richness which has been unpacked as measures of lexical variation (diversity), lexical sophistication and lexical density. Lexical density has been defined as the ratio of lexical words (e.g., nouns, verbs and adjectives) to grammatical function words (all other parts of speech (e.g., prepositions) (Read, 2000). The review that follows focuses deliberately on the constructs of diversity and sophistication because doubts have been raised as to the status of density as a measure of vocabulary (See Durrant et al., 2021) because arguably density equally tells us more about syntax than lexis and also because the measure of density has had no traceable influence on the study of formulaic language.

As the review of studies that measure the relationships between these measures and writing quality will show, the measures used are arguably not absolute measures of these types of knowledge and therefore it should be appreciated throughout the review presented here in Chapter Three and also in Chapter Four that what is measured is a fuzzy approximation of a writer's proficiency. This fuzzy approximation is guided by and to an extent restricted by (a) researchers' measure selection and rationale for such measures and (b) working with a snapshot of their production in the form of an often-one-time written text.

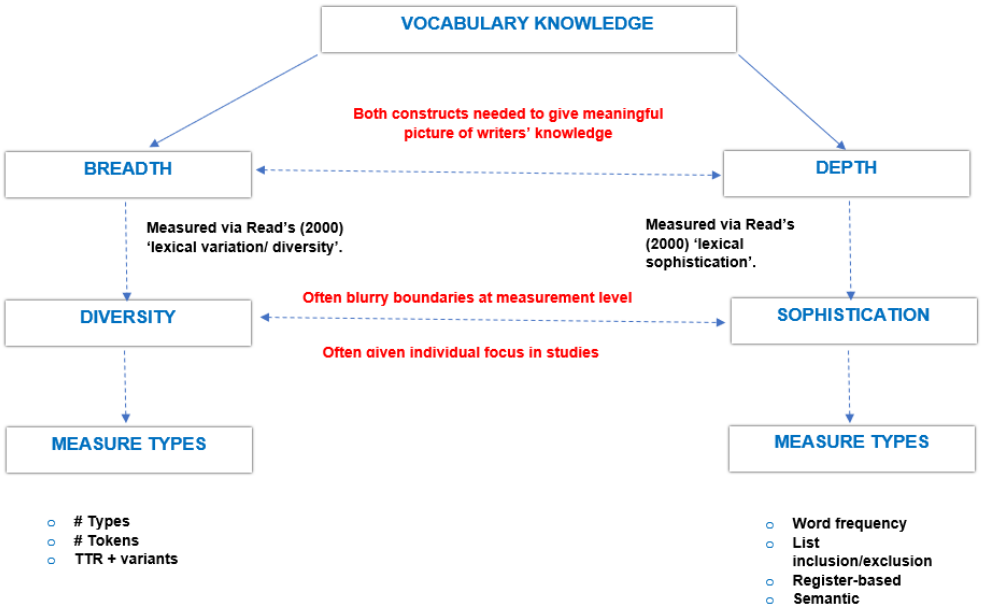


Figure 2: Map of Vocabulary Knowledge and Quantitative Measures

3.4 The Relationship between Vocabulary and Writing Quality

The study of writing quality and its relationship to linguistic features is concerned with examining learners' texts as products whereby the processes involved in producing the texts are given little to no attention (Polio & Friedman, 2016; Silva, 1993). This text-based work counts features according to various principles and under several different frameworks.

Although some researchers (e.g., Lu, 2012) have attached their study of diversity and sophistication to the term of 'lexical complexity', this term is not one which is mentioned in the vocabulary knowledge focused literature that is reviewed above and for many researchers (e.g., Kyle & Crossley, 2015), their guiding framework is still framed around notions of breadth and depth. For this

reason, the review that follows also continues to lean on this traditional vocabulary assessment literature[11].

Researchers have studied the complexity of learners' lexical systems (evidenced through producing written texts) under the construct of lexical complexity or lexical richness. Read (2000) unpacks 'lexical richness' to reveal that it operates under three main constructs: lexical diversity (lexical variation), lexical sophistication and lexical density. Lexical diversity is concerned with the amount of different words used by a writer and this can therefore be attached to the kind of vocabulary knowledge that 'breadth' is concerned with. This typically involves a distinction between all words (tokens) and words that are different (types). These distinctions have led researchers to count the number of types, the number of tokens and to create ratios that tap into variation or diversity by dividing the number of types by the number of tokens, in what is known as a Type-Token- Ratio (TTR) with multiple mathematical variants that attempt to minimize the effect that longer texts will naturally mean a lower type-token ratio because writers naturally use fewer new words as text length increases (Malvern et al., 2004). These trajectories are referred to in Figure 2 under diversity measures (referred to as # types, # tokens and TTR + variants). Sophistication is defined by Read (2000, p. 200) as the selection of low-frequency words that are appropriate to the topic and style of the writing, rather than just general, everyday vocabulary".

In tapping into diversity and sophistication, researchers have followed two broad pathways of choosing to either manually count diversity so manually

[11] It is also felt that this notion of lexical complexity is more grounded in L2 literature but for readers here it may be synonymous with lexical richness. Overviews of complexity and its measurement are presented in Ellis and Barkhuizen (2005), Gyllstad et al (2014), Housen et al., (2012), Lu (2012), Pallotti (2009, 2015), Wolfe-Quintero et al (1998).

counting types and tokens or parts of speech; or to count features by using automated computational tools that count types, tokens and parts of speech according to built in criteria and functions.

Researchers have included manual counts as their operationalisation (e.g. Hillocks, 2002; Taguchi, Crawford & Wetzel, 2013) where texts are physically read by the researchers and specific features are counted; however, these counts now typically rely on a combination of techniques from corpus and computational linguistics. These techniques and tools include taggers and parsers which are able to automatically identify individual words of a particular class and words sharing a syntactic relationship (Grant & Ginther, 2000; Bulté & Housen, 2014; Wolfe-Quintero, Inagaki & Kim,1998) and lexical and syntactic analysers that are able to count features before statistical analyses are carried out (Kyle & Crossley, 2015).

These techniques, often operating under the umbrella of Natural Language Processing (NLP), have facilitated the examination of diverse lexical features, including numerical counts of types and tokens in a text (e.g. Arthur, 1979; Becker, 2010; Bulté & Housen, 2014; Daller, Turlik & Weir, 2013; Douglas, 2015; Hou, Verspoor & Loerts, 2016; Kim, 2014; Treffers-Daller, Parslow & Williams, 2018; Vajjala, 2018; Vidakovic & Barker, 2010), as well as measures of sophisticated vocabulary.

In their study of diversity and sophistication, researchers have also used an amalgam of internal and external measures. Meara and Bell (2001) label internal measures as those which are frequency counts across the learner corpus under study while external measures involve comparing learner corpora frequencies, proportions and ranges with external counts from native reference

corpora and/or externally produced lists of academic vocabulary. It is this internal-external division that is used below to highlight the different research trajectories that have been used in studying writing quality. It should be noted that the review only includes measures and issues that are directly pertinent to informing the study of phraseology.

3.4.1 Diversity

Researchers across L1 and L2 contexts have measured vocabulary diversity in a number of ways ranging from early mathematically simple calculations to more recent mathematically complex calculations which use whole and segmented learner texts (e.g., Daller, Turlik & Weir, 2013; Douglas, 2015; Espada-Gustilo, 2011; Hou, Verspoor and Loerts, 2016; Kim, 2014; Kobayashi & Rinnert, 2013; Levitzky-Aviad, 2012; Levitzky-Aviad & Laufer, 2013; Reid & Findlay, 1986; Treffers-Daller, Parslow and Williams, 2018; Vajjala, 2018; Wang, 2014). At tertiary level, there is a lack of studies that focus on L1 writers with most research investigating L2 ESL and EFL learners. The review that follows therefore primarily draws on both L1 and L2 studies. A complete set of reviewed studies, their measures of diversity and findings are presented in Appendix B.

3.4.1.1 Basic TTR measures

At the simplest level, researchers have measured learners' diverse vocabulary by basic counts of internal measure types across texts or across a normalised frequency when text lengths differ (e.g. per 1,000 words) (e.g., Daller et al., 2013; Douglas, 2015; Hou et al., 2016; Kim, 2014). Studies have also counted tokens (Daller et al., 2013; Douglas, 2015) and used a Type-Token Ratio (TTR) which measures the number of types divided by the number of tokens to gauge diversity (Espada-Gustilo, 2011; Kim, 2014; Levitzky-Aviad, 2012; Levitzky-Aviad &

Laufer, 2013; Treffers-Daller, Parslow and Williams, 2018; Vajjala, 2018; Wang, 2014).

Irrespective of the measures used, researchers in these studies have compared learners' diversity with that of others in the same learner corpus. Overall, these studies found that lexical diversity increases significantly with proficiency level and that diverse vocabulary is a predictor of essay quality (Espada - Gustilo, 2011; Kim, 2014; Santos, Nerbonne & Verspoor, 2013; Treffers-Daller et al., 2018). In the early university texts from Espada-Gustilo's (2011) Filipino students, lexical diversity increased across proficiency levels and yielded a positive correlation to overall writing quality. Equally, in the early university essays from Kim's (2014) Korean students, lexical diversity increased with proficiency level and also correlated positively with writing quality. However, researchers have strived to provide more methodologically sound choices because TTR has been shown to be strongly influenced by text length. Text length influences the traditional TTR because as a text increases in length it naturally contains fewer new types and therefore the ratio naturally decreases as text length increases. This means TTR as a measure of diversity is influenced not by the writer's diverse vocabulary but by the natural decline in variation that longer texts experience (Malvern, Richards, Chipere & Duran, 2004).

This influence has led to several 'modified' alternatives being suggested namely: 'corrected TTR' (e.g., Arthur, 1979; Hou et al., 2016; Kim, 2014; Vajjala, 2018) and text segmented mean or standardised TTR (e.g., Becker, 2010; Grant & Ginther, 2000; Staples & Reppen, 2016; Vidakovic & Barker, 2010; Vajjala, 2018) as well as complex mathematical formulas such as 'D' (e.g., Bulté & Housen, 2014; Crossley & McNamara, 2012; Jarvis, 2002; Treffers-Daller et al., 2018; Yoon & Polio, 2017; Yoon, 2017); and 'MTLD' (e.g., Perin & Lauterbach,

2018; Riazi, 2016; Treffers-Daller et al., 2018; Vajalla, 2018) which are explained in Sections 3.4.1.2 and 3.4.1.3.

3.4.1.2 Modified TTRs

Carroll (1964) developed the 'Corrected TTR' (CTTR), which he calculated by dividing the number of word types by the square root of two times the total number of words. Arthur (1979) used Carroll's (1964) CTTR in his longitudinal study of 14 mixed L1 writers at university level however learners' diverse vocabulary was not a discriminatory feature between holistically graded essays when measured in this way. Other studies have found different in that CTTR has been able to discriminate between lower and higher quality essays in L2 adult exam-based contexts with Vajjala's (2018) corpora of argumentative essays, emails and letters indicating that diverse writing received higher CEFR and TOEFL grades irrespective of topic and L1 variables which were included in the regression models. Similarly, in a rare FYC study, Eckstein and Ferris' (2018) comparison of L1 and L2 undergraduate texts revealed L1 writers displayed more lexical variation in their texts than their L2 counterparts.

In an attempt to standardise counts, some researchers have counted types within a fixed text length ('Standardised Type-Token Ratios', or STTR). In L2 U.S university writing, Becker (2010), found a correlation between STTR and writing quality across samples of 200 words in U.S undergraduate level essays. A smaller group of researchers have used a sequence of 50 words yet students' diverse vocabulary was not a predictor of quality in tertiary level writing when measured in this way (Becker, 2010; Grant & Ginther, 2000; Hou et al., 2016; Kim, 2014) or in exam focused writing across CEFR-contexts (Vidakovic & Barker, 2010; Vajjala, 2018).

Another modification in the form of the Guiraud's Index (also known as Root TTR) accounts for text length by dividing the number of types by the square root of the number of tokens. Studies operationalising diversity in this way have reported significant increases across ESL and EFL university contexts (Bulté & Housen, 2014; Eckstein & Ferris, 2018; Hou et al., 2016; Kim, 2014). In their comparison of L1 and L2 FYC texts, Eckstein and Ferris (2018) found that L1 writers had significantly more varied lexis measured by the Root TTR than L2 writers.

A less common approach to diversity has focused on variation amongst lexical words only. This has been operationalised both as a simple count of lexical types (e.g., Celaya & Naves, 2009; Santos, Nerbonne & Verspoor, 2013), and as a TTR (e.g., Casanave, 1994; Engber, 1995; Laufer, 1994; Linnarud, 1986 & Nihalani, 1981) although their relationship to writing quality and proficiency level has varied across contexts. In her undergraduate level texts, Engber (1995) found diverse vocabulary as a lexical word TTR was a moderately significant correlate of writing quality with intermediate and advanced proficiency levels. However, no significant correlation with writing quality was found with Nihalani's (1981) Indian university writers when their homework essays were analysed.

Bringing these results together, when diversity is operationalised using these modified measures we see largely positive relationships between diversity and writing quality. In other words, these measures indicate that using more varied words appears to correlate positively with writing quality. The next sections explain how the operationalisation of diversity has been mathematically further developed in an attempt to minimize the influence of text length skewing the results.

3.4.1.3 Mathematically complex TTR modifications

More mathematically complex diversity measures have also tried to minimise text length issues. These include the Measure of Textual Lexical Diversity (MTLD) and D. In brief, MTLD checks each word in a text in sequence to see whether it has occurred before and calculates a ratio of total to unique words to each point, resetting when a pre-determined ratio is reached. D calculates the TTR in progressively larger portions of a text, generating a curve which traces the decrease in TTR as text length increases. The value D describes the shape of that curve, with higher values of D indicating greater overall diversity. MTLD shows a significant correlation with writing quality across L1 child and L2 adult writing (e.g., Aryadoust, 2016; Mazgutova & Kormos, 2015; Perin & Lauterbach, 2018; Riazi, 2016; Spurling, 2014; Treffers-Daller et al., 2018 & Vajjala, 2018). MTLD was a positive correlate of essay quality in a short EAP course in the UK in Mazgutgova and Kormos' (2015) study. This was also the case in the CEFR-based texts from Vajjala (2018). D also appears to positively correlate with writing quality (e.g., Aryadoust, 2016; Bulté & Housen, 2014; Crossley & McNamara, 2012; Daller et al., 2013; Knoch, Rouhshad & Storch, 2014; Qin & Uccelli, 2016; Treffers-Daller et al., 2018; Wang, 2014; Yoon & Polio, 2017; Yoon, 2017; Yu, 2010). In studies of U.S university writing, Bulté and Housen (2014) found D to be a significant correlate of writing quality while in the high-school writing of Chinese learners in Hong Kong, D was also found to positively correlate with writing quality when student letters and essays were graded holistically (e.g., See Crossley & McNamara, 2012; Qin & Uccelli, 2016).

The above section has reviewed individual word diversity and its relationship to writing quality across L1 and L2 writing. The review indicates a wealth of measures have been used to measure diversity with a focus on

rectifying text length issues. Drawing together this literature, students' use of diverse vocabulary can be said to have *some* relationship to writing quality with several studies indicating that diverse vocabulary was positively related to an increase in awarded grades.

Bringing together the results from using these mathematically advanced measures is more of a challenge because of the varied nature of the measures used and the different results obtained. However, overall, these measures seem to indicate that diverse writing has a positive correlation with writing quality. The next section considers how sophisticated vocabulary use has a relationship to and can predict grade score across a range of L1 and L2 contexts.

3.4.2 Sophistication

The last section showed a reasonably clear link between the theoretical construct of vocabulary breadth, the measure of lexical diversity and how this has been operationalised by counting the number of words in texts. The review now turns to consider the link between the theoretical construct of vocabulary depth and sophistication. As indicated in Figure 2, the definitions drawn on to tap into sophistication are not only grounded in depth but also mention breadth. As this review will show and as many researchers have indicated (e.g., Kyle & Crossley, 2015), many measures capture information about breadth and depth under the label of 'lexical sophistication'. This combination of knowledge types in relation to sophistication is made clear in the work of prominent measurement researchers Kyle and Crossley (2015, p.759) who, in their introduction to their automated tool TAALES (Tool for the Automatic Analysis of Lexical Sophistication) comment that although no agreed upon definition of sophistication exists, it generally involves the "depth and breadth of lexical knowledge available

to speakers, readers and writers" as specified in Meara (1996) and Read (1998). Therefore, it should be pointed out that in visualising vocabulary knowledge types and their relationship to breadth and depth that while diversity seems to be more clearly related only to breadth, sophistication is in contrast related to both breadth and depth and this comes across clearly in how researchers have gone on to create a number of objective quantitative measures that attempt to capture a property of either breath or depth in student writing. A full[12] quantitative measure list for sophistication is presented in Appendix B.

3.4.2.1 *Frequency-based wordlists: Lexical Frequency Profiles*

A popular operationalisation of sophistication has been the use of frequency bands. The most widely-used frequency band approach emanates from Laufer (1994) and Laufer and Nation (1995), who used Nation's (1984) frequency wordlists to determine sophistication. Laufer (1994) operationalised sophistication by using frequency wordlists divided into 1,000-word bands believing that as frequency decreases, words become more sophisticated and specialised in their use. Laufer and Nation (1995) developed the 'Lexical Frequency Profile' (LFP), which specifies the amount of vocabulary in a text taken from each band (e.g., Douglas, 2015; Laufer, 1998; Lemmouh, 2008; Levitzky-Aviad & Laufer,2013; Krzeminska - Adamek, 2016; Ruegg, Fritz & Holland, 2011; Verspoor, Schmid & Xu, 2012). Studies have found that learners used more words from the second and third 1,000 words frequency bands as proficiency level increases (Laufer, 1994; Vidakovic & Barker, 2010) and more words from the second and third 1,000-word frequency bands were used by writers scoring higher grades across university placement exams and course-work assignments

[12] Again, this is limited to the studies and measures that have gone onto to play some kind of guiding role in the study of phraseology.

(Douglas, 2015; Lemmouh, 2008). The underlying principle of vocabulary knowledge that researchers appear to trying to tap into could be related to the amount of low-frequency vocabulary that writers use and therefore how much they appear to show knowledge of extending use of 'everyday' vocabulary to use more lower-frequency words that appear to be more commonly found in academic writing. However, in many of these studies, this kind of link between theoretical construct or knowledge type is not made explicitly.

The Advanced Guiraud Index also measures sophistication by isolating 'advanced' types and dividing them by the square root of the number of tokens. 'Advanced' types are defined as those not appearing on West's (1953) 1st and 2nd 1,000 General Service wordlists (Daller, Van-Hout & Treffers-Daller, 2003). Sophistication measured in this way was not found to correlate with text quality in the 90 U.S university essays that Bulté and Housen (2014) analysed.

Under operationalisation from the LFPs, sophisticated vocabulary appears to be positively correlated with writing quality: better writers use more words from the wordlists. In line with the LFP approach, the use of the Advanced Guiraud index appears to also be an attempt to tap into the amount of words not found on general vocabulary lists and therefore implies that the more advanced types not found on the general list appear to have a relationship with higher quality writing implying this writing is more in line with the broad style of academic writing. However, also in line with the LFP approach, this measure's connection to a particular type of knowledge or construct of vocabulary is often not strongly stated.

3.4.2.2 *Register-based wordlists*

Another common approach to sophistication has looked at use of words from a specific register, on the assumption that use of such words is a marker of sophistication. A wide range of studies have used register-based lists but the most common register-based list is Coxhead's (2000) Academic Word List which is frequently embedded in the Lextutor tools from Cobb (2017) (e.g., Knoch et al., 2014; Knoch, Rouhshad, Oon & Storch, 2015; Qin & Uccelli, 2016; Storch, 2009; Storch & Tapper, 2009; Verspoor, Lowie, Chan & Vahtrick, 2017). In brief, these wordlist measures operate by counting in frequency or percentage how many words from a particular list a writer uses, or in some cases, how many words they use which are not on a list is taken to be a marker of sophisticated vocabulary use, depending on the nature of the list and the rationale set out by the researcher. The use of these register-based lists here appears to superficially tie in with breadth of vocabulary knowledge by tapping into the contexts that words are used in.

In their study of a single Dutch linguistics student, Verspoor et al (2017) found that the number of AWL words per text increased across the course time (4 years) and this number was also positively correlated with essay score. In another university level study, Storch and Tapper (2009) investigated the use of AWL words by postgraduate mixed L1 students in Australia. They found that the level of AWL words used increased as the course progressed, although no attempts were made to relate this relationship directly to quality score ratings. The ESOL-based work of Vidakovic and Barker (2010) also supports students' use of the AWL and its relationship to quality and proficiency constructs. In their

study of ESOL Skills for Life essay responses, they found that the number of AWL words used increased with proficiency levels (CEFR A2-B2 levels).

Other wordlists have also been studied less frequently with, Laufer (1994) looking at the University Word List longitudinally across a two-week period. In her study of 48 mixed L1 Freshman students' essays, she found that essays written two weeks apart differed in their frequency of UWL words with the later essay containing more on-list words. Again, like the LFP results, there appears to be support for a positive correlation with these register-based lists. Although it should be noted that there is less work using these register-based lists than traditional wordlists.

3.4.2.3 External cross-checks of native corpora

Sophistication has also been operationalised by cross-checking learner production with large-scale native corpora (Crossley, Salsbury, McNamara & Jarvis, 2010; Guo, Crossley & McNamara, 2013; Kyle, 2017; Kyle & Crossley, 2015, 2016; Mazgutova & Kormos, 2015; Perin & Lauterbach, 2018; Riazi, 2016). Measures that are based on checks with large-scale native corpora assign each word a frequency and/or range score - the former referring to the frequency of the word in the reference corpus and the latter referring to the number of texts in which it appears. The text is then given an overall score based on the mean of all the words that have been given a score. The underlying measure or theory behind such measures is that quantifying the amount of lower frequency words in a text acts a proxy for signalling sophisticated vocabulary use because low fewquency words are thought to be more sophisticated. Range measures are an attempt to counter the fact that a basic frequency count does not consider that individual texts may have very frequent or very infrequent occurrences of a word. That is to

say, a particular text may have high usage of a word, while another text may not have particularly high usage and so range measures take this imbalance into account. Range is also termed dispersion, entropy or contextual diversity with Kyle and Crossley (2016) explaining that words with a high range values occur widely across a number of different texts and contexts. Words with lower range values tend to more restricted in use to a smaller number of texts and contexts.

Different versions of these measures are sometimes given based on the choice of reference corpus, the types of words included in the counts (all words, content words, or function words) and whether a log transformation is applied to the scores (Kyle & Crossley, 2015). There is evidence to suggest sophistication operationalised in this way is task dependent, but few concrete assertions can be made as studies use different corpora that are based on different language registers and points in diachronic time (e.g. the SUBTLEXus corpus looks at frequencies of words found in the subtitles of American TV series while the CELEX frequencies are based on the psycholinguistic processing times of words). The choice of the CELEX corpus for some studies may be due to the fact that words with longer processing times may be thought of as being more sophisticated (Guo et al., 2013) while the use of the SUBTLEXus corpus may have been selected because it is based, not on formal language expected in academic writing, but largely colloquial/conversational domains, and can therefore show contrasts between high and low frequency words.

Using the automated tool Coh-metrix, Guo et al (2013) found that CELEX-based frequencies were highly significant indicators of quality for integrated TOEFL tasks but not for independent tasks written by US university students. Kyle and Crossley (2016) found that SUBTLEXus measures were not significant predictors in models of TOEFL tasks while Kyle's (2017) study of source-based

US university learner texts yielded weak correlations with quality for content and for all words SUBTLEXus frequencies.

These results may not be surprising since these corpora do not appear to be particularly representative of academic writing (e.g., the SUBTLEXus is based on subtitles of American TV shows and the CELEX is based on processing times, which may not be the most robust indicator of sophisticated words).

Taking these measures and results together, vocabulary that is less frequent and register specific in that it is found on a compiled list of register specific vocabulary has been found to correlate with writing grades across a range of assessment contexts. Sophisticated vocabulary, as identified by frequency and range-based measures in native reference corpora, also appears to correlate with writing quality.

The review also brings to the surface a number of important observations about the seemingly intertwined nature of breadth and depth or diversity and sophistication. The review highlights how in the case of those measures captured under sophistication, there continues to be a fixation with counting ‘how many’. This is clear in those measures termed frequency-based wordlist measures, where researchers continue to count ‘how many’ words are deemed sophisticated or not; in the case of those measures termed register-based wordlist measures, the same approach means there is still a focus on counting how many words appear to be connected to different registers. That is to say, the focus is still only on ‘ticking off’ presence on a list and therefore the level of knowledge that writers may be demonstrating with respect to depth is simply the ability to write the word. There is therefore only a shallow level of depth being tapped into, that is intertwined with that of breadth.

The review of diversity and sophistication also reveals a focus on methodological steps that attempt to find the 'best' measure of diversity and sophistication. However, despite these efforts, unsurprisingly, no one measure has yielded consistent reliable results. This is in part because research contexts vary significantly with different learner groups, language backgrounds, text and task types. Yet the above review shows that despite measurement differences, diverse and sophisticated vocabulary appears to have a relationship with writing quality and that researchers should specify their measure choice according to their available resources, literacy in using automatic tools and performing calculations, and research context.

3.5 Summary

This chapter's review of measures of diversity and sophistication appears to indicate that more diverse vocabulary use and vocabulary use that is less frequent and more register specific are both being markers of writing quality across a number of levels of writing contexts. The next chapter of the study details how the construct of collocation acts a kind of type of vocabulary knowledge and how the construct has been identified and measured across learner texts. In doing so, the chapter also draws attention to the gaps that exist in the recent study of collocation and describes how the present study addresses these gaps.

Chapter Four: Relationships between Collocations and Writing Quality

4.1 Introduction

This chapter serves as a literature review of recent research that has examined the relationship between collocations and writing quality. The chapter begins by situating the notion of collocation within the wider phenomena of phraseology and phraseological units. After defining phraseology in Section 4.2 and setting out the range of phraseological units and their extraction methods from language corpora, Section 4.3 narrows to focus on collocations, their composition, extraction and properties. Section 4.4 focuses on the importance of association measures which offer a way of understanding collocation. Section 4.5 underlines the importance of collocation complexity to writing quality in terms of how the diverse and sophisticated use of collocations has a relationship to writing quality. Section 4.6 sets out the range of measures that have been used to operationalise diverse and sophisticated collocation use and explicitly draws attention to the limitations of these measures, and their lack of consistency across contexts. Within this section, the chapter examines how the emerging work that this study aims to build on has uncovered a range of relationships between collocations and writing quality. Section 4.7 then highlights the role of task, language status and individual raters when determining the relationship between collocations and writing quality and finally Section 4.8 indicates how the study makes a contribution to knowledge by measuring this feature-quality relationship by taking into account the variables highlighted in Section 4.7.

4.2 Defining Formulaic Language

Formulaic language is the cover-all term that is used to describe particular word strings or word combinations (Wray, 2002) which vary in length (number of words) from a combination of two words to an indeterminate maximum, which according to early scholars can include sentence-length combinations (Jespersen, 1924/1976). According to Cowie (1994), formulaic or phraseological units are recognised or conventionalised ways of saying things. Cowie (1988) distinguishes between 'formulae', which perform pragmatic functions in texts such as greetings '*Good morning*' and '*how are you?*' and 'composites' which function syntactically below sentence level (e.g., '*dry run*', '*close shave*') and focus on creating meaning including units such as restricted collocations (e.g., '*blow a fuse*') and (figurative and pure) idioms (e.g., '*blow your own trumpet*' and '*blow the gaff*') which do not have an obvious functional purpose but have a more or less single meaning (Cowie, 1988; Granger & Paquot, 2008). These combinations also vary structurally by being composed of different word classes which include lexical and grammatical strings such as the lexical composite '*blow a fuse*' and the grammatical composite '*under the microscope*' whereby lexical composites include more than one lexical word and grammatical composites include more than one grammatical word.

However, the status of a word combination being 'formulaic' or 'phraseological' differs greatly depending on the theoretical position adopted and the vastly different frameworks used for identification (Wray, 2002). Wray (2002, 2008) notes the lack of consensus in defining phraseology with no fewer than 40 terms used to describe word combinations that appear formulaic. Amongst the most commonly used terms to encapsulate the nature of formulaic sequences are the following: chunks, collocations, composites, idioms, fixed expressions,

formulaic language, formulaic sequences, formulae, lexical phrases, prefabricated patterns and multi-word units (Wray, 2002). Myles and Cordier (2017) and Ellis (2008) highlight how formulaic sequences are labelled differently depending on the theories of language that are adopted. They note that formulaic sequences are labelled ‘chunks’ in psycho-linguistically-oriented or usage-based theories of language because learners are said to mentally ‘chunk’ combinations together and store these as wholes in long-term memory for later retrieval during communication, while corpus-linguists often label sequences as ‘clusters’ presumably due to extraction tools that extract words surrounding a search or node word in a cluster.

Despite differences in terminology, it *is* agreed that these word combinations consist of words that have especially strong relationships with each other (Wray, 2008). Wray (2002, p. 9) highlights the nature of a formulaic sequence as: “A sequence continuous or discontinuous of words or other elements, which is or appears to be, prefabricated: that is stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar”. This means language users may view these units, that comprise of several words, as single units which are retrieved from memory when needed. However, the belief about holistic storage and the research base that it is founded on has been criticised as being inconclusive because it has often relied on single method reaction time and processing-advantage type studies that only offer limited evidence as support for the beliefs (Siyanova-Chanturia, 2015; Wray, 2002).

4.3 Approaches to Identifying Formulas and their Units of Interest

In defining formulaic language, Wray (2002) highlights how formulaicity is not generally an absolute, concrete quality. Instead formulaicity can be arbitrary or subjective for a language user or a particular group of language users that share characteristics. Wray (2002) and Wood (2015) acknowledge that this subjectivity is likely to influence judgements from native and non-native speakers as to a particular combination's status. For example, when asked to judge or identify formulaic language, identification is known to vary widely across individual native speakers, and that their views on formulaicity varies over time and context. This observation also makes generalisability across research contexts problematic.

Despite this acknowledgement, a plethora of researchers have sought to provide guidance and frameworks on identifying and classifying formulaic language (e.g., Cowie, 1998; Howarth, 1998a, 1998b; Lewis, 1993, 1997; Nattinger & DeCarrico, 1992; Wray & Perkins, 2000; Wray, 2002). In their studies, researchers have separated these phenomena according to multiple criteria including but not limited to frequency, structural variability, pragmatic or discourse function and internal compositionality (Biber, Conrad & Cortes, 2004; Cowie, 1991, 1994, 1998; Howarth, 1996, 1998a, 1998b; Nesselhauf, 2005; Sinclair, 1991). These criteria help extract distinct types of phraseological units in terms of their internal and externally-decided features, identification methods and their respective overlaps.

These criteria can be broadly thought of or contained within two main paradigmatic lens that relate to linguistically-oriented views of language and how it may be learned: frequency-based approaches and phraseological-based approaches[13] (e.g.,, See Bestgen & Granger, 2014; Chen, 2019; Cowie (1981, 1994, 1998); Durrant & Schmitt, 2009; Gablasova, Brezina & McEnery, 2017; Granger & Bestgen, 2014; Granger & Paquot 2009; Gyllstad and Wolter 2015; Henriksen, 2013; Nesselhauf, 2005; Paquot (2018, 2019); Wray, 2002 for examples of these approaches in action). Since the present study draws on the frequency approach only, the sections that follow focus on this in depth.

4.3.1 Frequency-based Approaches

Two central guiding principles govern frequency-based approaches: recurrence and co-occurrence. Recurrence is the repetition of the same word strings either by a single language user or by a group of language users (Ellis, 2008; Granger & Paquot, 2009; Gries, 2008; Kjellmer, 1994). This recurrence is directly observable by looking at frequencies of word pairings that appear together in corpus data (Evert, 2004). Under a frequency-based school of thought, co-occurrence relates to the occurrence of words together in a deliberate manner and cannot be explained by chance alone (Evert, 2004). To clarify the different foci of recurrence and co-occurrence, Evert (2007) notes the mere repeated occurrence of word pairs is not a sufficient indicator for a strong attraction

[13] In brief, the phraseology approach can be summarised as relying on semantic criteria rather solely frequency of occurrence (Granger & Paquot, 2008). It is not within the boundaries of this study to set out such pros and cons of the frequency or phraseology approach. Instead readers are directed to summaries in Granger and Paquot (2008), Gablasova et al., (2017) and Nesselhauf (2005) which spell out the phraseology approach as an approach that takes into consideration the semantic relationship between words and their degree of non-compositionality of their meaning.

between the words, meaning that a word pair may be frequent but there may not be a strong attraction between the individual words, in other words a combination may be frequent but the words may be able to take many other partners.

In the majority of studies that have used frequency as an identification principle, these sequences are typically contiguous or adjacent combinations (words appear one after the other). A sole focus on recurrence means corpus tools such as *KfNgram* (Fletcher, 2002-2005), *Wordsmith* (Scott, 2018), *Antconc* (Anthony, 2018) and *Collocate* (Barlow, 2018) extract all units that meet a basic frequency threshold (e.g., 3 occurrences per text or in some cases 'X' occurrences per million words) and in most cases a basic length threshold (typically set at 2-4 words). These units may be semantically or syntactically complete (e.g., '*plastic surgery'*) or incomplete (e.g., '*of the'*) and are extracted irrespective of their structural or semantic make up or their psycholinguistic salience or pedagogical relevance (Paquot & Granger, 2012; O'Donnell, Römer & Ellis, 2013). These extracted units are termed '*ngrams*' (word strings of 'n' length) with 'bigrams' consisting of adjacent two-word units and 'trigrams' consisting of contiguous three-word units. Bigrams commonly extracted in ngram studies include: '*of a*', '*is the'*, '*to the'*, '*there is'*, '*on the'*, '*have to'*, '*to be'*, and '*it is*' (Bestgen & Granger, 2014). These 'ngrams' are typically further classified into more linguistically and pedagogically useful sub-divisions including lexical bundles, p-frames and statistical collocations which researchers have gone to great lengths to distinguish between.

The chapter now turns to consider these divisions in more detail.

4.3.1.1 Ngrams, lexical bundles and p-frames

When extracted from raw data and simply counted, ngrams contain a mixture of structurally incomplete and complete word strings which may or may not perform discourse or pragmatic functions, in other words be labelled 'lexical bundles'; may or may not be semi-fixed with a variable slot, labelled a 'p-frame' or 'collocational framework'; or be a meaning-making 'lexical' or 'grammatical' collocation depending on the theoretical classification adopted by the researcher (Ellis & Vlach-Simpson, 2009; Gablasova et al., 2017; Granger & Paquot, 2009). Ngrams are often left qualitatively unanalysed and simply counted as their frequencies alone can help determine genre differences (e.g., Römer, 2010; Tang & Cao, 2015) and quality ratings in texts (e.g., Bestgen and Granger, 2014; Crossley, Cai & McNamara, 2012; Granger & Bestgen, 2014; Kyle & Crossley, 2016). This crude count has also been used recently to train automated feedback and grading system as well as machine learning tools (e.g., Crossley et al.,2012; Crossley, Defore, Kyle, Dai & McNamara, 2013; Deane & Quinlan, 2010).

Lexical bundles are extracted from texts by using corpus-based tools such as *KfNgram* (Fletcher, 2002-2005) and *Wordsmith* (Scott, 2018) which operate under a frequency threshold (set to no less than 3 words occurring in 'N' texts (e.g., 5 texts) or occurring 'X' times per 1,000,000 million words (e.g., 20 times per million words) (Allen, 2010; Appel & Wood, 2016; Biber, Conrad & Cortes, 2004; Chen & Baker, 2010, 2016; Ruan, 2016; Staples, Egbert, Biber & McClair, 2013; Zheng, 2016). This frequency threshold is often arbitrarily set however the seminal bundle work of Biber et al., (2004) sets this frequency as occurring 40 times per 1 million words. Other researchers, such as Chen and Baker (2016)

and Appel and Wood (2016), who work with much smaller learner corpora, set their frequency thresholds much lower. Most studies have tended to focus on lexical bundles of 4 words in length because 3-word bundles are often structurally or semantically incomplete 4-word bundles (Appel and Wood, 2016). Hyland, cited in Appel and Wood (2016), also notes that 4-word bundles have more easily identifiable functions and perform a greater range of these functions than 3-word bundles. Lexical bundles are extracted and then often separated into different categories of discourse function.

Biber et al's (2004) pioneering work devised a functional classification system which separates these bundles into stance, referential and discourse bundles. Under this system, stance bundles highlight the writer's views, referential bundles refer to literature or the work of others and discourse bundles take on a text organising role. Examples of these functional bundles are provided in Table 4.

Table 4: Types of Lexical Bundles

Bundle Type / Structure	Example	Representative Studies
Discourse	On the other hand The second theory is	Allen (2010) Staples et al (2013)
Referential	The role of the There are a lot	Durrant (2017) Staples et al (2013)
Stance	It is important to I would like to	Biber et al (2004) Staples et al (2013)

Researchers such as Chen and Baker (2016) have also separated these bundles into structural units of verb-phrase, noun-phrase and prepositional-based bundles. Examples of these structural bundles are provided in Table 5.

Table 5: Structurally-classified Lexical Bundles

Bundle Type / Structure	Example	Representative studies
Noun-phrase bundles	More and more people The purpose of the	Chen & Baker (2016) Ruan (2016)
Verb-phrase bundles	Is one of the It is difficult to	Chen & Baker (2016) Ruan (2016)
Preposition-based bundles	At the same time In terms of the	Chen & Baker (2016) Ruan (2016)

While easily extracted from raw data, Ädel and Erman (2012) recognise that bundles may overlap in their functions and so classification needs careful monitoring. They also emphasise that since these bundles are neither idiomatic in meaning, in the sense of fixed-meaning idioms (e.g., '*blow the gaff*'), or perceptually salient, they can be pedagogically limited in their value and researchers and practitioners need to choose bundle candidates that are suitable for language learning carefully.

P-frames are bundles that are fixed except for a variable slot in the sequence (Garner, 2016; Granger & Paquot, 2009; O'Donnell et al., 2013; Römer, 2010). P-frames help researchers examine variable sequences and are determined by software such as the suite of tools from Anthony (2018). Like lexical bundles, p-frames can be classified into functional frames and structures (Garner, 2016). Thus far, p-frame variability has helped distinguish between genre types (e.g., Römer, 2010) but their distribution, frequency and type in the study of writing quality has yielded weak results (Garner, 2016; O'Donnell et al., 2013). Examples of p-frame patterns are provided in Table 6.

Table 6:Types of P-frames

P-frame structural type	Example	Representative study
Verb-based frames	I * like to	Garner (2016)
Function - word frames	The * of the	Garner (2016)
Content-word frames	the * stage of	Garner (2016)

Note: (* = variable slot in the p-frame)

It should be clear from the presentation and discussion of ngrams, lexical bundles and p-frames that two important points emerge. The extraction of such units does not appear to well-grounded in tapping into a specific theoretical construct of language per se (Durrant et al., 2021). That is to say, what is being extracted is simply stretches of language that are frequently occurring but not always theoretically relevant or useful for language learning or assessment purposes. The added layer of theoretical use is added later by the researcher who determines from these extensive lists of extractions, units which may in fact be performing a particular textual function. In contrast, focus on the notion of collocation has a long-standing theoretical history dating back in corpus linguistics circles to the work of Firth (1957) and slightly earlier in the 1950s in informing work from document and symbol retrieval work (e.g., see the early work from Osgood (1952) and Osgood et al (1957), outlined in Stevens et al., (1964)). The remainder of the chapter focuses on the frequency-based approach taken to collocation.

4.3.2 Frequency-derived Collocations

Many scholars have pointed out that the phenomena of collocation has no agreed upon universal definition (e.g., Pecina, 2010) but under a frequency-based or distributional perspective, researchers have drawn on a number of guiding principles and definitions. Founding frequency-advocate Firth (1957), and later 'Neo-Firthian' scholars who have built on Firth's work such as Palmer

(1933/1966), Sinclair (1991, 1998), Halliday (1966) and Halliday and Hasan (2001) have determined collocations by working with the following definitions and characteristics. Firth (1957, p. 181) initially set out that: “Collocations of a given word are statements of the habitual or customary places of that word” whereby these habitual occurrences are frequently occurring in natural speech. Sinclair (1991) and Halliday (1966) advance Firth’s observations with Sinclair (1991, p. 170) defining collocations as: “the occurrence of two or more words within a short space of each other in a text”. Under a frequency or distributional perspective, co-occurrence of words is based on observing recurring words that appear together in a segmented or distance-based span (Evert, 2004; Jones & Sinclair, 1974).

Several other scholars who make use of the frequency approach also draw on the stances and definitions provided by Firth (1957) and Sinclair (1991). Evert (2004) comments that collocations are generally understood to be word combinations which are not completely predictable only from the basis of syntactic rules, and they should be listed in a lexicon and learned in the same way single words are learned. McIntosh, Francis and Poole (2009, p. 6) also consider a collocation as “a pair or group of words that are often used together” involving “the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance”. This definition therefore considers adjacent or contiguous and non-contiguous groupings as collocations as do Halliday and Hasan (2001, p. 317) who also state that one word is typically associated with another as they tend to occur in similar environments or contexts. The explicit focus for collocation-oriented researchers is on the syntagmatic relationship that words have. This syntagmatic relationship is concerned with how

words combine together and the arbitrary restrictions placed on this combinatory relationship (Carter & McCarthy, 1988; Granger, 1998; Schmitt & Schmitt, 2020).

Firth (1968, p. 181) believes that collocation is a type of 'mutual expectancy' between words. In this sense, collocating words predict one another, where we see one, we expect to see the other. Cruse (1986, p.40) also echoes the beliefs of Firth (1968) by stating: "In collocation, the constituent elements are, to varying degrees, mutually selective". This selection is referred to by Sinclair (1991, p. 173) as 'mutual choice'. These 'choices' have been said to be representations of an individual's psycholinguistic language makeup or system and are expressed by Sinclair (1987, p. 319) through the idiom or slot and filler principle. This principle is explained as a language user having a large number of semi-preconstructed phrases at their disposal. These phrases are single choices, even though it may appear that they can be analysable into segments. These seminal thoughts can also be traced back to the work of Osgood (1952) and Osgood, Suci and Tannenbaum (1957), summarised in Stevens et al (1964). Osgood (1952,pp.54-55) acknowledges in a lengthy quote that:

> "If in the past experience of the source, events A and B…have occurred together, the subsequent occurrence of one of them should be a condition facilitating the occurrence of the other: the writing or speaking of one should tend to call forth thinking about and hence producing the other"

This quote from Osgood (1952) also ties into the notion of tendency and typicality that Seretan (2011) summarises.

Seretan (2011) brings together a number of later definitions of collocation. For example, Hausmann (1985) defines a collocation as a typical, specific and characteristic combination of two words. Benson (1990) taps into the notion of collocation as an arbitrary and recurrent word combination while Smadja (1993) draws on recurrence and cooccurrence by stating that a collocation is a recurrent combination of words that co-occur more often than expected by chance and that correspond to arbitrary word usages. Bartsch (2004) also shares similar views as she considers collocations as "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other". These definitions all involve mention of frequency or recurrence and 'typicality' with Seretan (2011) pointing out that these terms are all equally familiar terms in statistics, implying that researchers interested in collocation can unpack these concepts through the use of statistics.

Under these collective views then, a collocation therefore has two important properties that this study adopts in defining and operationalising collocation:

- It comprises of two or more words that appear near each other.
- The appearance of these words together is recurrent and the words co-occur more often than could be explained by random chance.

The next sections of the review focus on how collocations have been identified and how cooccurrence as non-random chance has been operationalised statistically via measures of association.

4.3.2.1 *Positional and Relational Approaches to Cooccurrence*

Under Jones and Sinclair's (1974) views, collocations need not be contiguous so long as the words occur in a span of the search or node word. This span length has been debated however Jones and Sinclair's (1974) optimal window is 4 words left or right of the node[14]. This span or 'window' identifies collocations according to frequency irrespective of the syntactic relationship between the words (Nesselhauf, 2005). Jones and Sinclair's (1974) specification of this span has advanced initial observations from Firth (1957) who did not specify space limitations or what would constitute a space that resulted in collocational bonds being broken (Wood, 2015). Examples of this span approach are presented in Durrant & Brenchley (in press):

(1) *The old **dream** of wireless communication through space has now been **realized.***

(2) *She **realizes** that the buzzing sound from her **dream** is present in her bedroom.*

Evert (2004) and Durrant and Brenchley (in press) note that this span approach extracts pairings with different or no syntactic relations. In response to this the relational approach extracts combinations that have a syntactic relation or dependency. The extracted units are therefore those that fulfil a particular syntactic relationship that the researcher wants to explore (e.g., adjectives with a modifying dependency on a common noun) and are often extracted automatically with the help of a dependency parser (e.g., the Stanford parser, (Manning et al., 2014) extracts adjective modifying dependencies with a common noun as 'amod'

[14] Smadja (1993, p. 151) comments that this window is 5 words while Guiliano (1965) recommends a window of fixed length 7 words for certain kinds of text punctuation and states that sentence boundaries can be ignored when looking at contiguity association between words.

dependencies). These might include the modifying verb dependencies such as the noun + verb as direct/indirect object combinations (‘*He won the lottery’)* that Paquot (2018, 2019) has examined.

Seretan (2011) has also picked up on this point by explaining that in example (3), span based approaches are not always able to capture collocations because they are not in the vicinity of the span set:

(3) “*The **problem** is therefore clearly a deeply rooted one and cannot be **solved** without concerted action by all parties*”

Seretan (2011) also helps highlight how researchers can take advantage of using relational extraction by pointing out that span approaches also extract syntactic noise. In example (4) it is possible to extract “human organisations” when the only relations are “human rights” and “human rights organisations”.

(4) “*human rights organisations*”

The definition of collocation that Sinclair (1974,1991) taps into advances our understanding of collocation to include word pairings that have a statistical association or combinatory relationship. These ‘*statistical collocations*’ are based on frequency distributions and the probability of word 1 taking specific ‘partners’ as word 2 options. The most commonly referred to measures of this statistical relationship in the literature are measures of association which are divided into measures of significance and measures of the strength of association (Evert, 2004; Pecina, 2005, 2010).

The next section of the review provides an overview of these measures, their formulae and their applications in studies. The section concludes with an explanation of how the narrative surrounding these measures is often described

as ‘fragmented’ or essentially patchy and how this leads to gaps in obtaining a coherent narrative around these measures.

4.4 Measures of Association

Both Sinclair’s (1974,1991) comments that the individual words that appear in word combinations do so more often than chance would predict and Smadja’s (1993, p. 143) views that: “natural languages are full of collocations, recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages” offer a window into understanding (a) whether or not there is a statistically significant association between the words in the combination and (b) allows researchers to consider the strength of association in more detail.

In the case of point (a), there is an interest in how much confidence researchers can have in stating that these word combinations do not simply appear as a result of random chance and so tests whether there is a significant chance of occurrence that is not down to random chance. In the case of point (b), the degree of association is tapped into by considering that some word combinations have a higher degree of combinatory strength or ‘glue’ than others. Points (a) and (b) are captured statistically through the use of association measures. Evert (2004, p. 75) defines an association measure as “ a formula that computes an association score from the frequency information in a pair type’s contingency table”. The score is intended to be an indicator of how strong the evidence is that there is an association between the words in the word combination. In principle, high positive scores indicates there is strong evidence of association while high negative scores or scores close to 0 indicate weak evidence of association.

The contingency table calculates in the case of (a) the chance of this occurrence appearing to be the result of chance and in the case of (b) the strength of association between the words. An example of a basic contingency table is shown in Table 7 while how the contingency table is used to calculate expected frequencies measures is presented in Table 8.

Many scholars (e.g., Xiao, 2015; Evert, 2004; Manning & Schütze, 1999) have set out how a basic table operates by looking at the occurrences of the collocation candidate (u, v) where u and v are typically lemmas; by also looking at the values in the table for '¬ u' which represents any lexical term except 'u' and '¬ v' represents any lexical item except v. These respective occurrences are symbolised by the letter a – d and can be explained as follows for the observed frequencies in Table 7.

- 'a' is referred to as the joint frequency which is the number of occurrences for the collocation candidate.
- 'b' is the number of times that 'u' appears as word 1 but any other lexical item appears as word 2 (e.g., the occurrence of ¬ v).
- 'c' is the frequency of 'v' without the appearance of 'u'.
- 'd' is represented by the formula: N – (a + b) – (a – c) + a and is the corpus size minus the frequency of the word pair and the frequency of 'u' and ¬ v (a +b), minus the frequency of the collocation and the frequency of 'v' without the appearance of 'u' (a – c). In other words this is the total number of words remaining after deducting the frequencies involving either the word combination or individual word frequencies of either of its individual words.

- N is the total number of words in the corpus.
- R_1 (Row 1) is represented as the sum of a + b (the sum of all the pairs with v as the second word). This is referred to as the marginal frequency.
- C_1 (Column 1) is represented as the sum of a + c (the frequency of all pairs with v in the second position). This is is referred to as the marginal frequency.
- R_2 (Row 2) is represented as the sum of c + d. This is therefore the frequency information for word pairs with one word of the 'u v' combination (in this case the frequency of 'v' with another word that is not 'u' – depicted as ¬ u, plus the frequency information of d. This is referred to as the marginal frequency.
- C_2 (Column 2) is represented as the sum of b + d. This is therefore the frequency information for word pairs with one word of the 'u v' combination (in this case 'u' with another word that is not 'v' – depicted as ¬ v. This is referred to as the marginal frequency.

The row and column sum totals therefore reflect calculations that represent aspects of 'everything else' other than the frequency of the 'u v' combination, accounting for all other probabilities or frequency combinations.

Table 7: Example Contingency Table for Observed Frequencies

	Y = v	**Y = ¬ v**	
X = u	a	b	R_1 = a + b
X = ¬ u	c	d	R_2 = c + d
	C_1 = a + c	C_2 = b + d	N= a+b+c+d

Note: 'a' also = O_{11}; 'b' also = O_{12}; 'c' also = O_{21}; d also = O_{22} in some calculations.

For Table 8, the values of a – d represent the expected frequencies and are calculated as follows.

- ‘a’ is represented as the sum total of R_1C_1 / N.
- ‘b’ is represented as the sum total of R_1C_2 / N.
- ‘c’ is represented as the sum total of R_2C_1 / N.
- ‘d’ is represented as the sum total of R_2C_2 / N.

Table 8: Example Contingency Table for Expected Frequencies

	Y = v	**Y = ¬ v**	
X = u	a	b	R_1 = a + b
X = ¬ u	c	d	R_2 = c + d
	C_1 = a + c	C_2 = b + d	N= a+b+c+d

‘a’ also = E_{11}; ‘b’ also = E_{12}; ‘c’ also = E_{21}; d also = E_{22} in some calculations.

Points (a) and (b) can be looked at as two paths with which researchers approach the nature of cooccurrence and subsequently collocation. This has led researchers such as Evert (2004) and Pecina (2005,2010) to use these paths as a way of grouping together different kinds of measures. Evert (2004) distinguishes between measures that test statistical **significance of association** (discussed in Section 4.4.1) and measures that test the **strength of association** (discussed in Section 4.4.2). Evert (2004) also adds on a number of measures as **information theory** derived measures (discussed in Section 4.4.3) and **heuristic measures** (discussed in Section 4.4.4) that can be blends of both groups of measures. The sections that follow draw on these groupings to explore the range of associations that have influenced linguistics research and also second language learning and assessment research (discussed in Section 4.6).

In doing so, the review draws on theoretical support from the guiding work of Evert (2004) and Pecina (2005, 2010) as well as studies from across the corpus and computational linguistic divide.

4.4.1 Significance of Association

The first batch of association measures that are covered in this section involve some kind of statistical test of significance with the most common measures falling under the label of 'hypothesis testing measures'. Evert (2004) breaks this large group down into three separate groups: likelihood measures, exact hypothesis measures and asymptotic hypothesis measures. The measures that fall under the 'significance of association' label aim to quantify the amount of evidence that the observed sample provides against the non-association of a given pair type (i.e. deciding whether or not to accept or reject the hypothesis that there is an association between the words).

In the first category of these significance measures, Evert (2004) presents a number of *likelihood measures* that are set out in Table 9. These measures calculate the probability of the observed contingency table. Evert's (2004) narrative highlights the evolution of the likelihood measures as each measure 'builds' on the previous by phasing out the incidence of mathematical bias in that each measure pays attention to skewed frequencies and therefore favours some types of word combinations over others.

Table 9: Likelihood Measures

Measure Grouping	Measure	(Simplified) formula*	Representative studies
Likelihood measure	Multinominal likelihood	$\frac{N!}{N^N} \cdot \frac{(E_{11})O^{11} \cdot (E_{12})O^{12} \cdot (E_{21})O^{21} \cdot (E_{22})O^{22}}{O_{11!} \cdot O_{12!} \cdot O_{21!} \cdot O_{22!}}$	Wiechmann (2008) Pecina (2005, 2010)
	Hypergeometric likelihood	$e^{-E_{11}} \frac{(E_{11})O_{11}}{O_{11}!}$	Wiechmann (2008) Pecina (2005, 2010)
	Binomial likelihood	$\frac{N!}{N^N} \cdot \frac{(E_{11})O_{11} \cdot (E_{12})O_{12} \cdot (E_{21})O_{21} \cdot (E_{22})O_{22}}{O_{11}! \cdot O_{12}! \cdot O_{21}! \cdot O_{22}!}$	Wiechmann (2008) Pecina (2005, 2010)
	Poisson-likelihood	$e^{-E_{11}} \frac{(E_{11})O^{11}}{O_{11!}}$	Wiechmann (2008) Pecina (2005, 2010)
	Poisson-Stirling log measure	$O_{11} \cdot (\log O_{11} - \log E_{11} - 1)$	Wiechmann (2008) Pecina (2005, 2010)

In the second category under significance measures, *exact hypothesis tests* calculate the significance or p-value of the observed frequency data. These include the computationally costly to calculate Fisher's exact test with this measure being favoured over the log-likelihood measures by Pedersen (1996).

In the next category of association measures, the *asymptotic statistical tests* calculate a test statistic that can be translated into an approximate p-value. These include measures presented in Table 10. Seretan (2011) explains that the hypothesis tests looks to see if a hypothesis about the population is supported by evidence data and in the language expressed in Tables 7 and 8, the hypothesis tests whether the items u and v in the candidate pair 'u,v' are dependent on each other or not, i.e. if they are in fact independent. This means that the test statistic indicates how far the observed frequency deviates from what would be expected under the null hypothesis.

These measures have been extensively covered by scholars such as Manning and Schütze (1999) and form the backbone of the kind of statistical measures that are included within corpus platforms such as AntConc and Wordsmith (Xiao, 2015). Evert (2004) distinguishes between measures such as the t-score and the chi-squared with those that use test statistics in a different way: likelihood ratio tests. These are ratios between the maximum likelihood of the observed data under the null hypothesis and its unconstrained maximum likelihood. They are listed as the log-likelihood ratio, the log-likelihood, and the log-likelihood ratio (Dunning) with the latter from Dunning (1993).

Table 10: Asymptotic Hypothesis Testing Measures

Measure family	Hypothesis testing measure	Formula	Representative studies
Statistical tests of independence	t-score	$\frac{O_{11-}E_{11}}{\sqrt{O_{11}}}$	Pecina (2005) Wiechmann (2008)
	z-score	$\frac{O_{11-}E_{11}}{\sqrt{E_{11}}}$	Berry-Rogghe (1973) Smadja (1993)
	Chi-squared test (X^2)	$\frac{N\ (O_{11-}\ E_{11})^2}{E_{11}\ E_{22}}$	Krenn (2000) Evert & Krenn (2001)
	Fisher's exact test	$\frac{f\ (x*)!\,f(\bar{x}*)!\,f(*y)!\,f(*\bar{y})\bar{!}}{N!\,f(xy)!\,f(x\bar{y})!\,f(\bar{x}y)!}$	Pecina (2005, 2010)
	Log likelihood ratio (Dunning)	$-2\ \log\frac{L(O_{11,}C_{1,}\,r)\ \cdot L\,(O_{12},C_{2,}r)}{L(O_{11},C_{1,}r_1)\cdot\ L(O_{12}C_2r_2)}$ $L(k,n,r)=r^k\ (1-k)1^{n-k}$ $r=\frac{R_1}{N},r_1=\frac{O_{11}}{C_1},r_2\ =\ \frac{O_{12}}{C_2}$	Dunning (1993) Evert (2004)
	Loglikelihood[a]	$2\ \sum_{ij} O_{IJ}\ \log\frac{O_{ij}}{E_{ij}}$	Evert (2004)
	Squared log likelihood ratio[a]	$-2\sum_{ij} f_{ij\ \log\frac{f_{ij}}{\bar{f}_{iJ}}}{}^2$	Pecina (2005, 2010)
	Poison significance measures	$\frac{f(xy)-f(xy)\log f(xy)+\log f(xy)!}{\log N}$	Pecina (2005, 2010)

Note: [a] denotes a measure classified slightly differently in Pecina (2005) or another reference study.

At this point in the chapter, it is important to recognise that many frequencies and formulae have been presented, and it seems helpful to clarify to readers how these formulae operate in an actual word pair calculation. To illustrate these formulae 'in action', the calculation of the t-score for the word pair 'sweet smell' from Xiao (2015) is set out. The t-score is used here as a representative illustration of a hypothesis based measure.

Xiao (2015)[15] walks readers through an example of how the t-score would be calculated for 'sweet smell' by using frequencies from the BNC. Xiao (2015, p.110-111) sets out that there are:

- 90 occurrences of 'sweet smell' (represented as a in contingency table).
- 3,460 occurrences of the word 'sweet'
- 3,508 occurrences of 'smell'.
- 98,313,429 total words in the corpus.

Under the t-score formula in Table 10: $\frac{O_{11}- E_{11}}{\sqrt{O_{11}}}$, the t-score would be calculated as O_{11} = a, E_{11} = a expected. To calculate E_{11}, we need to know 'b', 'c' and 'd'.

- 'b' is calculated as the number of occurrences of word 1 ('sweet') minus the number of occurrences of 'sweet smell' ('a'). This means b = 3460 – 90 = 3,370. .
- 'c' is calculated as the number of occurrences of word 2 ('smell') minus the number of occurrences of 'sweet smell' ('a'). This means c = 3508 – 90 = 3418.

[15] Other examples of contingency tables can be found in Evert (2004), Seretan (2011) and Xiao (2015).

- ‘d’ is calculated as N – the number of word 1 occurrences – the number of word 2 occurrences + number of occurrences of the collocation candidate; meaning d = 98,313,429 – 3,460 – 3,508 + 90.
- E_{11} is therefore calculated as R_1C_1 = a+b a+c / N so = 90 + 3,370 * 90 + 3418 / 98,313,429 = 0.123.

The t-score is therefore calculated as $\frac{O_{11} - E_{11}}{\sqrt{O_{11}}} = \frac{90 - 0.123}{\sqrt{90}}$ = 9.474.

The t-score has been interpreted in many language studies as being worthy of consideration when it is above 2. A score above 2 is considered to be indicative of a collocation. A greater score is therefore indicative of being a stronger collocation candidate with high scoring t-scores also comprise of words that are particularly high frequency and therefore appear in many domains, registers and disciplines (Durrant & Schmitt, 2009).

4.4.2 Degrees of Association

Similar to those measures labelled as significance measures, Evert (2004) also makes divisions between groups under the ‘degree of association’ category. Table 11 shows the first of these as point estimates. Evert (2004) introduces this group under the advantage that unlike the significance measures, they do not have bias in favour high frequency items. Evert (2004) compares a number of measures here by pointing out the relationships between mutual expectation, Jaccard and the Dice coefficient which are mathematical transformations of each other. Evert (2004) also notes how the ‘Geometric mean’ is the square root of the MI^2. Evert (2004) also comments that although minimum sensitivity has not

been widely used in experiments for extracting collocations it has performed well when used.

Table 11: Point Estimates

Measure family	Point estimate measure	(Simplified) formula*	Representative studies
Point estimates	MI	$\log \frac{O_{11}}{E_{11}}$	Evert (2004) Pecina (2005)
	Relative risk	$\log\left(\frac{O_{11}C_2}{O_{12}C_1}\right)$	Evert (2004) Pecina (2005)
	Dice coefficient	$\frac{2O_{11}}{R_1\ C_1}$	Evert (2004) Pecina (2005)
	Mutual expectation	$\frac{2f(xy)}{f\ (x*)+f(*y)} \cdot P(xy)$	Evert (2004) Pecina (2005)
	Jaccard[a]	$\frac{a}{a+b+c}$	Evert (2004) Pecina (2005)
	Geometric mean	$\frac{O_{11}}{\sqrt{R_1C_1}} = \frac{O_{11}}{\sqrt{N}E_{11}}$	Evert (2004)
	Minimum sensitivity	$\left\{\frac{O_{11}}{R_1} \cdot \frac{O_{11}}{C_1}\right\}$	Evert (2004)
	Odds ratio	$\frac{ad}{bc}$	Evert (2004)

In relation to this grouping, other work with an interest in utilizing many measures has been that of Pecina (2005, 2010) and Pecina and Schlesinger (2006). Pecina' s work has also largely followed the groupings from Evert (2004) however, Pecina's work introduces a number of coefficient measures which do not feature in Evert (2004) or are only briefly mentioned. These coefficients and their formulae are discussed in Table 12. The goal of Pecina's exploratory work has been to statistically evaluate how successful these measures are at extracting 'true' collocations as later judged by human evaluations. It may therefore be unsurprising that the distinct theoretical and mathematical differences between these measures are not expanded on in detail. However, a comment to be made here is that the measures in Table 12 range from the

mathematically simple (e.g., the odds ratio) to the mathematically complex (e.g., the Gini index). Table 12 also highlights how some measures are an amalgam of others (e.g., the Klosgen measure uses the ‘Added Value’ measure in its calculation).

Table 12: Other Coefficient Measures from Pecina (2005, 2010)

Coefficient measure	**(Simplified) formula***
Russel-Rao	$\frac{a}{a+b+c+d}$
Sokal-Michiner	$\frac{a+d}{a+b+c+d}$
Rogers-Tanimoto	$\frac{a+d}{a+2b+2c+d}$
Hamann	$\frac{(a+d)-(b+c)}{a+b+c+d}$
Third-Sokal Sneath	$\frac{b+c}{a+d}$
First Kulczynsky	$\frac{a}{b+c}$
Second Sokal Sneath	$\frac{a}{a+2(b+c)}$
Second Kulczynsky	$\frac{1}{2}\left(\frac{a}{a+b}\right)+\left(\frac{a}{a+c}\right)$
Fourth Sokal Sneath	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$
Odds ratio	$\frac{ad}{bc}$

Yulle's ω	$\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$
Yulle's Q	$\frac{ad - bc}{ad + bc}$
Driver-Kroeber	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Fifth Sokal Sneath	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Pearson	$\frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Baroni-Urbani	$\frac{a + \sqrt{ad}}{a + b + c + \sqrt{ad}}$
Braun-Blanquet	$\frac{a}{\max(a+b, a+c)}$
Simpson	$\frac{a}{\min(a+b, a+c)}$
Michael	$\frac{4(ad - bc)}{(a+d)^2 + (b+c)^2}$
Mountford	$\frac{2a}{2bc + ab + ac}$
Fager	$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2}\max(b, c)$
Unigram subtuples	$\log\frac{ad}{bc} - -.3.29\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

U cost	$\log\left(1+\frac{\min(b,c)+a}{\max(b,c)+a}\right)$
S cost	$\log\left(\frac{1+\min(b,c)}{a+1}\right)-\frac{1}{2}$
R cost	$\log\left(1+\frac{a}{a+b}\right)\cdot\log\left(1+\frac{a}{a+c}\right)$
T combined cost	$\sqrt{U\ x\ S\ xR}$
Phi	$\frac{P(xy)-(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*)(1-p(*y)}}$
Kappa	$\frac{P(xy)+P(\bar{x}y)-P(x*)p(*y)-P(\bar{x}*)P(*\overline{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\overline{y})}$
J measure	$\max\left[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(\bar{x}y)\log\frac{P(\overline{y}\|x)}{P(*\overline{y})},\ \ P(xy)\log\frac{P(x\|y)}{P(x*)}\right.$ $\left.+P(\overline{x}y)\log\frac{P(\bar{x}\ y)}{P\ (\overline{x})}\right]$
Gini index	$\max[P(x*)(P(y\|x)^2+P(\overline{y}\|x)\text{^}2\ -P(*y)\text{^}2+P(\bar{x}*)\ (P(y\|\bar{x})^2+$ $P(\bar{y}\|\bar{x})^2-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2+P(*\bar{y})(P(x\|\bar{y})^2+P\left(\overline{x\|\bar{y}}\right)^2$ $-\ P(\bar{x}*)^2]$
Confidence	$\max[P(y\|x),P(x\|y)]$
LaPlace	$\max\frac{N\ P(xy)+1}{N\ P(x*)+2},\frac{N\ P(xy)+1}{N\ P(*y)+2}$
Conviction	$\max\left[\frac{P\ (x*)P(*y)}{P(x\bar{y}\ )},\frac{N\ P(\bar{x}*)P(*y)}{P(\bar{x}y)}\right]$
Platersky-Shapiro	$P(xy)-P(x*)P(*y)$

Certainty Factor	$\max\left[\frac{P(y\|x)-P(*y)}{1-P(*y)}, \frac{P(x\|y)-P(x*)}{1-P(x*)}\right]$
Added Value	$\max[P(y\|x)-P(*y), P(x\|y)-P(x*)]$
Collective Strength	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)} \cdot \frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P\bar{x}\bar{y})}$
Klosgen	$\sqrt{P(xy)\cdot AV}$

Note: The use of P(xy) etc can be replaced by the same a-d letters as other formulas have included in their representation.

4.4.3 Information Theory: Mutual Information

Evert (2004) introduces the next category of measures as being grounded in concepts of entropy, cross-entropy and mutual information and so are labelled 'information theory-derived' measures. These measures summarised in Table 13 quantify the non-homogeneity of the observed contingency table compared to the contingency table of expected frequencies. Most of these measures centre around Mutual Information (MI) measures. The MI score is a non-directional measure of mutual expectancy/attraction where words A and B are said to attract each other's presence equally (Evert, 2004; Gablasova et al., 2017). Mutual Information is explained by Evert (2004) as expressing the 'overlap' between two events or distributions. Pointwise MI compares two events 'A' and 'B' and is the logarithmic ratio of their actual observed joint probability to the expected joint probability if A and B were independent, as shown by the formula in Table 13. As shown in Table 13, many other variants of the Pointwise MI exist including the average MI and local MI however these have featured sporadically in the literature (e.g., they are not mentioned in Pecina (2005, 2010)).

The MI score is a normalised score that is comparable across corpora however there is no theoretical maximum or minimum with researchers seemingly setting their own arbitrary cut-off points for collocational status (Gablasova et al., 2017). This has typically been an MI score of 3 or more for a two-word pairing and anything lower than 3 is taken to be non-collocational in nature and has so far been labelled as 'less interesting' for this reason (Church & Hanks, 1990).

Some of the measures discussed so far are recognisable and are included in publicly available corpus platforms and tools. For example, Xiao (2015) notes that AntConc includes the MI and t-score, while more modern tools such as LancsBox (version 5.4) able to calculate these collocation measures (Brezina, Weill-Tessier., & McEnery, 2020).

Table 13: Information Theory Measures from Evert (2004) and Pecina (2005)

Measure family	Common association measure	(Simplified) formula*
Mutual information, its derived measures and other measures from information theory	Pointwise MI	$\log \frac{P\ (xy)}{P\ (x\ *)P(*\ y)}$
	Average MI	$\sum_{ij} O_{ij}\ \cdot\ \log \frac{O_{11}}{E_{11}}$
	Local MI	$O_{11}\ \cdot\ \log \frac{O_{11}}{E_{11}}$
	Mutual dependency	$\frac{\log P\ (xy^2)}{P\ (x\ *)P\ (y\ *)}$
	Log frequency biased mutual dependency	$\log \frac{P(xy)^2}{P\ (x\ *)P(y\ *)} + \log P(xy)$
	Normalised expectation	$\frac{2f\ (xy)}{f\ (x\ *) + f(y\ *)}$
	Mutual expectation	$\frac{2f(xy)}{f\ (x\ *) + f(*\ y)}\ \cdot P(xy)$
	Salience	$\log \frac{P(xy)^2}{P\ (x*)P(*y)} \cdot \log f(xy)$

Another group of measures that can be traced back to Pecina (2005, 2010) are labelled ‘context’ measures as shown in Table 14. Although the exploratory literature has not extensively described these context measures, they essentially involve taking into account the context surrounding the candidate pair. Pecina (2010, p. 143) gives an example where the context measures take into account the context either side of the word and factors this into equation. These measures would therefore appear to be computationally costly.

Table 14: Context Measures from Pecina (2005, 2010)

Context measure	(Simplified) formula*
Context entropy	$-\sum_w P(w\,\vert C_{xy})\log P(w\vert C_{xy})$
Left context entropy	$-\sum_w P(w\,\vert {C^l}_{xy})\log P(w\vert {C^l}_{xy})$
Right context entropy	$-\sum_w P(w\,\vert {C^r}_{xy})\log P(w\vert {C^r}_{xy})$
Left context divergence	$P(x*)\log P(x*) - \sum_w P(w\,\vert {C^l}_{xy})\log P(w\vert {C^l}_{xy})$
Right context divergence	$P(x*)\log P(x*) - \sum_w P(w\,\vert {C^r}_{xy})\log P(w\vert {C^r}_{xy})$
Cross entropy	$-\sum_w P(w\,\vert C_x)\log P(w\vert C_y)$
Reverse cross entropy	$-\sum_w P(w\,\vert C_y)\log P(w\vert C_x)$
Intersection measure	$\frac{2\vert C_x \cap C_y\vert}{\vert C_x\vert + \vert C_y\vert}$
Euclidean norm	$\sqrt{\sum_w (P(w\vert C_x) - P(w\vert C_y))^2}$
Cosine norm	$\frac{\sum_w P(w\vert C_x) - P(w\vert C_y))}{\sum_w P(w\vert C_x)^2 \cdot \sum_w P(w\vert C_y)^2}$
LI norm	$\vert P(w\vert C_x)\, P(w\vert C_y)\vert$
Confusion probability	$\sum_w \frac{P(x\vert C_w)P(y\vert C_w)P(w)}{P(x*)}$
Reverse confusion probability	$\sum_w \frac{P(y\vert C_w)P(x\vert C_w)P(w)}{P(*y)}$
Jensen-Shannon divergence	$\frac{1}{2}\left[D\left(P(w\vert C_x)\,\vert\vert\,\frac{1}{2}(P(w\vert C_x) + P(w\vert C_y)\right) + D\left(P(w\vert C_y)\vert\vert\,\frac{1}{2}(P(w\vert C_x) + P(w\vert C_y))\right)\right]$
Cosine of pointwise MI	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\,\sqrt{\sum_w MI(w,y)^2}}$

KL divergence	$\sum_w P(w\|C_X) \log \frac{P(w\|C_x)}{P(w\|C_y)}$
Reverse KL divergence	$\sum_w P(w\|C_y) \log \frac{P(w\|C_y)}{P(w\|C_x)}$
Skew divergence	$D(p(w\|C_x) \,\|\| \infty (w\|C_y) + (1-\infty) p\,(w\|C_x))$
Reverse skew divergence	$D(p(w\|C_y) \,\|\| \infty (w\|C_x) + (1-\infty) p\,(w\|C_y))$
Phrase word cooccurrence	$\frac{1}{2}\left(\frac{f(x\|C_{xy}}{f(xy)} + \frac{f(y\|C_{xy})}{f(xy)}\right)$
Word association	$\frac{1}{2}\,(\frac{f(x\|C_y)-f(xy)}{f\,(xy)} + \frac{f\,(y\|C_x)-f(xy)}{f\,(xy)})$
Cosine context similarity	$\frac{1}{2}(\cos(C_x\, C_{xy}) + \cos(C_{y,}\, C_{xy}))$ $C_z = (z_i); \cos(C_{x\,,} C_y) = \frac{\sum x_i y_i}{\sqrt{\sum x_{i^2} \cdot \sqrt{\sum y_{i^2}}}}$
In Boolean vector space	$z_i = \delta\,(f(w_i \mid C_z)$
In tf vector space	$z_i = f(w_i \mid C_z)$
In tf idf vector space	$z_i = f(w_i \mid C_z) \cdot \frac{N}{df(w_i)}; df(w_i) = \|\{x\colon w_{i\,\in C_x}\}\|$
Dice context similarity	$\frac{1}{2}\,(dice\,(C_{x,} C_{xy}) + dice(C_{y,} C_{xy})$ $C_z = (z_i); dice\,(C_{x,} C_y) = \frac{2\,\sum x_i y_i}{\sum x_{i^2 + \sum y i^2}}$

4.4.4 Heuristic Measures

Evert's (2004) heuristic measures are the final category mentioned in his work. These measures are all either modified versions of previously introduced measures or they are measures that do not quite fit the previous groupings as shown in Table 15. Under Evert's (2004) classification, these measures also include MI^2, MI^3 and a measure which combines the MI and t-score together from Church et al (1991). Church et al (1991) combine the MI and t-score to produce the MI/t-score measure however this measure has not been widely adopted in the literature. Church et al (1991) ranked collocation candidates according to their association strength, measured by the MI, then they only retained those that also had significant evidence of association from the t-score results.

Table 15: Heuristic Measures

Measure family	**Common association measure**	**(Simplified) formula***	**Representative studies**
Mutual Information-derivatives	MI^2	$\log \frac{(O_{11})^2}{E_{11}}$	Evert (2004)
	MI^3	$\log \frac{(O_{11})^3}{E_{11}}$	Evert (2004)
Combined MI and t-score measure	MI/t-score	$\min \left\{ h_1 \left(log \frac{O_{11}}{E_{11}} \right),\ h_2 \left(\frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \right) \right\}$	Church and Hanks (1991)

4.4.5 Asymmetrical Measures

The hypothesis measures and common association measures outlined in previous sections are underpinned by the belief that the cooccurrence of words appears to be symmetrical in that the words in the pairings are mutually attracted to each other with neither word 1 or word 2 being responsible for attracting the other, in other words, the attraction between words is mutual. However, the linguistic reality is that this is an oversimplification and attraction between words is actually asymmetrical in that one word in the pairing is more strongly attracted to the other rather than the attraction being mutual. This asymmetrical attraction has more recently gained attention through the use of the Delta P measure. This measure does not feature in either Evert (2004) or the work of Pecina (2005, 2010). Although the principle of directionality is later commented on by Evert (2007), this shift from their original groupings seems to indicate a wider consideration for many other measures of association. Table 16 presents the formulas for the Delta P and some of the studies that have commented on its use as an association measure.

The Delta P w1 | w2 measure looks at the extent word 1 attracts word 2 while the Delta P w2 | w1 looks at the extent word 2 attracts word 1 and therefore appreciates that in language words are not symmetrical (Gries, 2013a; Schneider, 2018). Its values range from -1 to +1 with Schneider (2018) highlighting that it punishes word pairs whose second word also frequently occurs in other combinations.

Table 16: Asymmetrical Measures

Measure family	Common association measure	(Simplified) formula*	Representative studies
Asymmetrical measures	Delta P w1| w2	$\frac{a}{a+b} - \frac{c}{c+d}$	Schneider (2018) Gries (2013a)
	Delta P w2 |w1	$\frac{a}{a+c} - \frac{b}{b+d}$	Schneider (2018) Gries (2013a)

4.4.6 Comparing Measure Groups and Exploring Gaps in the Narrative

One of the difficulties of bringing this literature together is that the measures of association originate in different fields. Further to this, beyond the large-scale explorations of Evert (2004) and Pecina (2005, 2010), other association measure work has focused on a single measure or group (e.g., Dunning, 1993). These measures have also had the research ethos that there is one “best” measure that detects collocations. For these reasons, Gries and Durrant (2020) highlight the fragmented nature of the association measure literature and call for further exploration of a wider range of measures that appear in the literature. Wiechmann (2008) and Gablasova et al (2017) also call for researchers to better understand association measures.

Wiechmann (2008, p.267) highlights how researchers need a better understanding of how the range of measures are related to each other and Gablasova et al (2017) also encourage measure selection to be less arbitrary and more principled. A comment to be made here is that this understanding how measures are related remains an ongoing challenge because of the lack of narrative for some measures. In Evert (2004) and Pecina (2005, 2010), although measure families are set out, it is not always clear how closely related members of the families are to each other and it is also less clear how families related to others. Schneider (2018) comments that high collinearity between association

measures is unlikely however he finds the Delta P to be highly correlated with the transitional probability. Schneider (2018) comments that this finding is unusual however it does not seem so surprising when we consider that many association measures have been developed by individual researchers and are based on mathematically modifying the same basic information contained within the contingency tables of observed and expected frequencies.

The review has now defined collocation under a frequency-based approach and outlined how collocations have been identified and examined under this approach. The review now turns to examine how this frequency-approach has been drawn on by learner language researchers and how the properties of recurrence and co-occurrence have been studied as part of the focus on the complexity of collocation use by learners.

4.5 The Importance of Collocations in Student Writing

The phenomena of phraseology and its relationship to writing quality appears to be an area of scholarship that is, at present, almost exclusively focused on second language learners (cf. Durrant & Brenchley, in press). This may be due to the fact that the development of multi-word units such as collocations and discourse managing bundles is a vital component of native language use, a requirement for acceptance into discourse communities, and a marker of fluent language production (Wray, 2002). Collocations have been established as a particularly troublesome area for L2 learners as they struggle to master the often-arbitrary combinations, semantic constraints and non-literal fixed meanings of these combinations (Granger & Paquot, 2009). The lack of understanding of these arbitrary restrictions leads to L2 writers producing combinations that violate

these restrictions such as '*powerful coffee*' instead of '*strong coffee*' (Nesselhauf, 2005).

In this respect, several scholars have highlighted how learners who have the ability to make appropriate choices in their use of language have been shown to score highly across writing assessments (Crossley, Salsbury & McNamara, 2015; Granger & Bestgen, 2014). This ability also extends to the use of collocations and phraseology more widely where the mastery of formulas is an important contribution to language proficiency which allows writers to appear fluent and able to fulfil the basic communicative purpose that their texts strive for (Henriksen, 2013; Myles, 2004). Widdowson, cited by Wray (2002), notes that communicative competence involves knowing a range of ready-made patterns, formulaic frameworks and having a toolbox of these ready to meet the communicative needs of the speaker/writer and listener/reader.

Lewis (2000) also elaborates on the importance of these units by emphasising that since academic writing is chiefly focused on the accurate communication of complex information, (which is often knowledge shared by colleagues with a common background), common words, phrases and collocations are not only an essential facilitator of this complex information but are an expected occurrence when readers are fellow members of the same discourse or discipline community. The selection and use of phraseology across disciplines remains a constant challenge for learners where the use of collocations and particularly the aforementioned arbitrary restrictions that they entail is seen as a "kind of threshold" to specialised discourse at undergraduate level (Ward, cited in Marco, 2011).

Ellis (2008) also notes how each genre and discipline has its own phraseology and effective mastery of genres and inclusion into a discipline is dependent on mastering its relevant discourse therefore the use of collocations is an attempt to convey a particular meaning and show membership of a particular discourse community (Li & Schmitt, 2009; Myles, 2012). Bartholomae (1986, p. 146) also recognises the difficulties that students face when entering new discourse communities and explains how: “A large part of the undergraduate’s task, is thus to gain skills in using formulas to reflect an insider’s familiarity with the discipline specific discourse as well as to reflect the more general conventions we associate with academic English”. Wray (2002) elaborates on this skill in that language users make language choices that they know depict certain values, styles and groups of other language users. Wray (2002) continues to stress that it is the appropriate use of these choices that determines if a message is communicated successfully and if a writer gains acceptance into their intended discourse or disciplinary community.

These notions of choice making link to the now widely acclaimed notion of ‘native-like selection’ which has been championed by Pawley and Syder (1983). Pawley and Syder (1983, p. 194) view the use and mastery of formulaic language as a crucial marker of language proficiency and a necessity if learners are to attain ‘native-like’ competency in the language. Pawley and Syder (1983) explain that learners need to be able to know what word combinations are well-formed and which are unnatural or marked uses. These marked uses have particular importance for writing quality as they have been shown to be processed slowly by readers and also yield negative correlations to writing quality scores meaning that as their frequency increases, writing quality scores decrease (e.g., Bestgen & Granger, 2014; Crossley et al., 2012; Granger & Bestgen, 2014). In many

previous studies, these marked uses are often below threshold or absent combinations in large native corpora and could be considered erroneous (Granger & Bestgen, 2014).

However, despite the fact that this 'native-like' selection ethos has been the foundational starting point for many corpus linguists to base their views and subsequent analyses on, it can and has been argued that this view creates an overly simplistic impression of native writing being presented as a worthy 'model' that non-native speakers should aspire to because their writing is viewed as deficient when compared to native speaker linguistic feature profiles. In this respect, Römer (2009) is among a limited, albeit growing number of corpus and rhetoric scholars, who point out that L1 post-secondary writers are not blessed with academic writing ability and their mastery of such skills including appropriate genre or disciplinary phraseology should not be taken for granted.

Therefore in appreciating this acknowledgement from Römer (2009) and taking heed of the more 'equal' starting point that FYC scholars have presented in Chapter 2 (Sections 2.4 – 2.6) in that both L1 and L2 writers are new to university academic writing and both face challenges in adapting to academic writing, its conventions and expected language use, this study does not approach its inquiry with the assumption that L1 writers will have 'favourable' or 'superior' collocation use per se and instead bases its approach on an understanding that both groups will approach language use and writing tasks differently owing to the respective challenges they face.

The review now moves on to consider how researchers have measured the relationship between collocations and writing quality.

4.6 Complexity Measures and their Relationship with Writing Quality

The ability of combine words with their appropriate collocates or partners has been thought of as part of vocabulary knowledge under ‘depth’ (e.g., See Chapter Three, Section 3.3). However, we can equally conceptualise that the wide range of aspects of vocabulary knowledge can be applied to the notion of collocation itself (Gyllstad, 2007). That is to say, we can apply the same broad thoughts from vocabulary to collocation. We can think of these word combinations in terms of breadth (*how many* collocations do writers know/use?) and depth (what do they know about these collocations in terms of their syntactic and semantic makeup?; and how they are typically used in different registers and disciplines etc?). This underlying framework of knowledge is summarised in Figure 3. Since the focus of this study is specifically on collocation, the review of student writing that follows is solely focused on this type of phraseology. However, a complete picture of how diverse and sophisticated phraseology has been measured and the results from individual studies is presented in Appendix C.

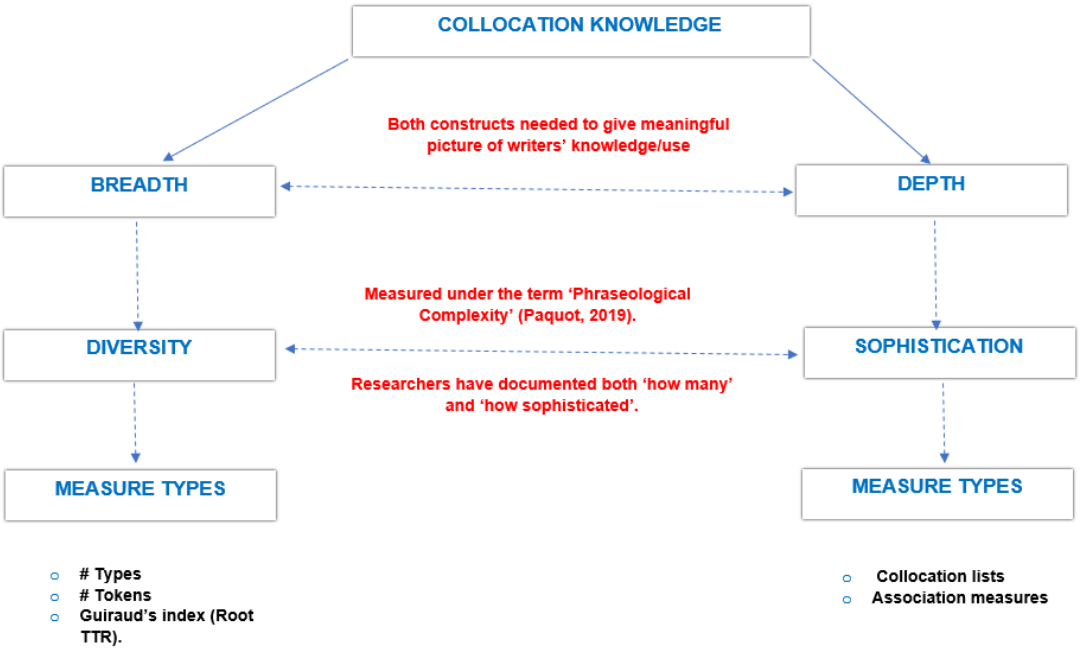


Figure 3: Map of Collocation Knowledge and Quantitative Measures

Despite the fact that the kind of framework or 'map' shown in Figure 3 has not been made explicit in studies of collocation-grade research, Paquot (2019) is one such study that does make this distinction. She draws on the same connections between breadth and depth and diversity and sophistication to ultimately coin the term "phraseological complexity". This term easily applies to collocation and the other types of word combination identified in Section 4.3.1; however, she uses the term specifically to study collocation. Phraseological complexity is defined as "the range of phraseological units that surface in language production and the degree of sophistication of such units" (Paquot, 2019, p.124). In her operationalisation of diversity and sophistication she taps into the former by calculating the Guiraud's index (Root TTR) as an equivalent analogy to one of the diversity measures used in single word vocabulary work (See Chapter Three, Section 3.4.1.2); and she taps into sophistication in two

respective ways. First, sophisticated collocations are taken to be those that appear in the Academic Collocation List from Ackermann and Chen (2013) (again making an analogy with the list-based approaches outlined In Chapter Three, Section 3.4.2.2). Second, sophisticated collocations are taken to be those that have particularly high MI scores with the assumption being that higher scoring MI combinations are indicative of more exclusive specialised word combinations that are particularly found in academic and discipline/ genre specific writing and ultimately considered to be beyond 'everyday' combinations. It should be pointed out here that, like vocabulary studies, there has been a focus on 'how many' also when it comes to tapping into sophistication.

4.6.1 Collocation Diversity

The only directly relevant study to the focus on collocation here is the diversity approach taken by Paquot (2019). Paquot (2019) remains the only researcher to date to explicitly focus on the diversity of word combinations by using syntactic dependencies akin to the types of dependency relations outlined in Section 4.3. In a corpus of texts from French L1 postgraduate linguistics students, she calculated the Root TTR for the following dependency types: amod (adjective modifier: 'She has black (JJ) hair (NN)'), advmod (adverb modifier: 'She has very (RB) black (JJ) hair'.) and dobj (direct object modifier: 'He won (VV) the lottery. (NN)*'*). Using this measure of diversity, she found no differences across proficiency levels. It is important to indicate here that this approach to diversity by Paquot (2019) includes all dependency combinations irrespective of their status as a collocation. It is simply a measure of the variation in dependency types. The next section of the review covers how researchers have studied collocation sophistication in different writing contexts.

4.6.2 Collocation Sophistication

In drawing parallels with the work on vocabulary, many researchers have used similar list-based approaches and measures that also appear in single word vocabulary work. This has included looking at sophistication by quantifying attested and non-attested combinations in a reference corpus (e.g., Bestgen, 2017; Bestgen & Granger, 2014; Granger & Bestgen, 2014; Garner et al., 2019; Kim et al., 2018; Kyle & Crossley, 2016; Paquot, 2018, 2019). It has also included looking at presence in a formula list (e.g., Paquot, 2018) or used a small number of the association measures outlined in Section 4.4 to study the relationship between collocations and writing quality (e.g., Bestgen, 2017; Chen, 2019; Granger & Bestgen, 2014; Garner et al., 2018, 2019; Kim et al., 2018; Paquot, 2018, 2019; Yoon, 2016). These approaches are discussed in the following sections.

4.6.2.1 *Cross-checks with Native Corpora*

In using external measures to operationalise sophistication, several studies have checked learner texts for the presence or absence of formulaic language that appears in large-scale native corpora (e.g., Bestgen & Granger, 2014; Bestgen, 2017; Garner et al., 2019; Granger & Bestgen, 2014; Kim et al., 2018; Kyle & Crossley, 2016; Paquot, 2018). Granger and Bestgen (2014) and Paquot (2019) focus on particular syntactic combinations which appear more than five times in a reference corpus. Granger and Bestgen (2014) looked at premodifier noun sequences (MN) which comprised of directly adjacent word pairs that were noun premodifier by another noun (labelled: NN) (e.g., ‘ozone layer’, or an adjective (labelled :JN). They also studied adjective modifying an adverb (labelled: AJ) (e.g., ‘incurably ill’). Across these combinations, they count the percentage of

combinations that appear less than five times in the BNC. They term these combinations ‘below threshold’. They found that these adjective + noun and adverb + adjective below threshold combinations were more common in advanced proficiency texts rather than intermediate texts. Paquot (2018) studied amod, advmod and dobj combinations by extracting these syntactic dependencies with the Stanford parser (Manning et al., 2014). Paquot (2018) found no significant differences between texts rated from B2 to C2 on the CEFR scale in terms of their use of these below threshold combinations.

Another take on the attested/non-attested approach has been to focus on combinations that are entirely absent from the reference corpus. Two studies have focused on these combinations. In their analysis of US university writing, Bestgen and Granger (2014) found that the proportion of types that were absent from COCA had a significant negative correlation with overall writing quality and language and vocabulary scores awarded to essays. Bestgen’s (2017) study of FCE and ICLE texts found similar. The proportion of absent types had a significant negative correlation with writing quality for FCE texts but a non-significant positive correlation for ICLE texts.

The underlying property or construct that this attested/non-attested approach aims to quantify is the degree of nativeness in learner texts with relationships largely indicating here that the more language found in the native reference source, the more it has a positive relationship with writing quality scores. A similar approach to quantifying this proxy of ‘nativeness’ has also been taken with the use of formula lists.

4.6.2.2 *Formula Lists*

Paquot's (2018) study used presence or absence in the Academic Collocation List (Ackermann & Chen, 2013) to distinguish collocation use between CEFR B2-C2 levels. In this particular study, presence or absence in the collocation list as a way of quantifying the amount of recognised native collocations that learners are using did not reveal strong significant findings. Across the broad operationalisations of attested and non-attested work, some interesting observations about the relationships between collocation and writing quality have emerged however more consistent clearer patterns of relationship have tended to be found when sophistication has been captured via association measures.

4.6.2.3 Association Measures

Several studies have used a small number of association measures from Section 4.4 to look at the relationship between collocations and writing quality (e.g., Bestgen, 2017; Granger & Bestgen, 2014; Garner et al., 2018, 2019; Kim et al., 2018; Paquot, 2018, 2019). Across the body of literature, researchers have focused on four association measures: MI, t-score, Delta P and Collexeme strength quantifying sophistication by mean scores and/or threshold scores. As Section 4.4 identified, the MI has a focus on the degree of exclusivity pairings may have and the measure's mathematical makeup means it award high scores to low frequency pairings, with high scoring MI combinations found to have narrower discipline specific uses (e.g., Durrant & Schmitt, 2009). A number of studies (e.g., Bestgen, 2017; Bestgen & Granger, 2014; Garner et al., 2018, 2019; Kim et al., 2018) have therefore used the Mean MI with bigram combinations and found a significant positive correlation between Mean MI and

writing quality. Across dependency types, Paquot (2018, 2019) also found the same significant positive relationships between Mean MI and writing quality.

Much of the work on threshold MI groupings has been influenced by Durrant and Schmitt (2009) who allocated bands to the MI scores. They choose between 7 bands from MI = 3-3.99 to MI= ≥ 10 whereas Granger and Bestgen (2014) and Paquot (2018, 2019) condense these categories into high, medium, low and non-collocational with MI scores of ≥ 7 deemed high MI collocations and those deemed low collocations those with MI scores of 3-4.99, and those < 3 deemed non-collocations. High-scoring MI combinations include instances of: '*conscientious objectors*', '*technological advances*', '*nitrous oxide*', '*densely populated*', '*bated breath*' and '*preconceived notions*' (Granger & Bestgen, 2014; Durrant & Schmitt, 2009) whereas low-scoring MI combinations are more generic in use and comprise of words that are less exclusively attracted to each other (Granger & Bestgen, 2014).

In an analysis of native and non-native writing, Durrant and Schmitt (2009) manually identified pre-modifying adjective-noun bigrams and examined both MI and t-scores to discover that native writers used more higher scoring MI combinations than non-native writers in university essays and that non-native writers tended to use more high scoring t-score combinations. Based on this observation, Durrant and Schmitt (2009) help emphasise that non-native writers tend to overuse more generic combinations while avoiding the use of more specialised combinations that would be more appropriate for disciplinary/academic writing.

Granger and Bestgen (2014) and Paquot (2018, 2019) build on this work by Durrant and Schmitt (2009). In their study of thresholds, Granger and Bestgen (2014) found that advanced level writers used more high scoring MI collocation types and tokens than intermediate level writers with intermediate level writers instead using more non-collocation types. Paquot (2018) also builds on this work by studying amod, advmod and dobj dependencies with French L1 postgraduate student texts. These results present a less clear-cut picture of student writers' dependency use. When dependencies were classified into low, medium and high MI scores, there was a significant increase in medium MI premodifier noun and verb direct object dependencies. There was also a significant increase in low MI premodifier noun dependencies but a significant decrease in premodifier noun non collocations. Results here are difficult to synthesise perhaps because of their different writing contexts but also because they have extracted combinations in different ways.

The Mean MI yields a clearer interpretable picture over the research landscape. Studies using the Mean MI on its own have largely found significant positive correlations with writing quality across exam contexts (e.g., Bestgen, 2017; Garner et al., 2018), across U.S university writing (Bestgen & Granger, 2014) and coursework writing (e.g., Paquot, 2018, 2019). Unlike the more varied picture for threshold MI results, there seems to be much more consistent evidence that holds across contexts that the Mean MI is indeed a significant correlate and predictor of writing quality.

A more fine-grained use of the MI has been to take a Mean MI per POS. This was the approach taken by Paquot (2018, 2019). Paquot found a statistically significant increase between B2-C2 CEFR levels for amod dependencies however B2-C1 and C1-C2 CEFR levels did not have significant increases. For

the Mean MI advmod dependencies (Paquot, 2018, 2019) found that there was a significant increase across CEFR levels but this was only between B2-C1 and B2-C2 levels and not between C1-C2 levels. For the Mean MI dobj dependencies, Paquot (2018,2019) found a significant increase only between B2-C2 and C1-C2 CEFR levels. These results highlight that fine-grained differences exist for each dependency type.

Another association measure studied in terms of thresholds is the t-score. Granger and Bestgen (2014) consider a t-score ≥10 as high, ≥ 6 and < 10 as medium and ≥ 2 and < 6 as low, and < 2 considered non-collocation. Granger and Bestgen (2014) found that intermediate writers used more high-scoring t-score combinations than advanced writers. They found that instead advanced writers used more medium and low scoring t-score combinations. When split for POS, intermediate writers used more high-scoring premodifier + noun combinations, more adjective + noun combinations and more adverb and adjective combinations. In contrast, advanced writers used more high scoring noun + noun combinations. For medium t-score combinations, no statistically significant patterns appear across intermediate and advanced proficiency levels. The same pattern of non-significant difference also appears for low-scoring t-score POS combinations. In terms of non-collocations, intermediate writers used significantly more premodifier + noun combinations and adjective + noun combinations.

Bestgen and Granger (2014) and Bestgen (2017) both use the Mean t-score in their analyses. Bestgen and Granger (2014) found in US university writing that there was a weak non-significant correlation with overall writing quality and language and vocabulary scores for types and tokens. In a study of FCE and ICLE writing, Bestgen (2017) found a significant positive correlation with FCE

writing quality but a non-significant positive correlation with ICLE text quality. Garner et al (2018) found that there was a significant increase across CEFR levels A2-B2.

Two other measures that have been less frequently studied are the Delta P (see overviews in Gries, 2013a; Schneider, 2018) and Collexeme Strength. Garner et al (2018) found in Korean L1 writing that the mean Delta P from COCA academic significantly increase across CEFR A2-B2 levels and had a significant positive correlation with writing proficiency levels. The Mean Delta P calculated from the spoken component of COCA also significantly increased across A2-B2 CEFR levels. A final measure that has only featured in Garner et al (2018) has been the Collexeme Strength measure. In this study, the mean Collexeme Strength based on the COCA academic was found to significantly increase across A2-B2 CEFR levels.

Given the increasing number of ways that sophisticated collocations have been operationalised, it is perhaps unsurprising that a clear, coherent picture of their relationship with writing quality is more difficult to ascertain. However, the review of such measures robustly illuminates association measures as measures that are consistently related to writing quality grade scores across various contexts.

4.6.3 Research Gap I: Association Measure Selection

At this point in the review it is necessary to recognise that these association measure studies allow us to develop an understanding of how the different association measures, with their different formulas, illuminate different properties and uses of language from student writers. However, it should also be noted that

the student writing literature that has used measures of association have overwhelmingly used the MI and t-score measures. While, this choice may seem natural if they want to examine particular exclusive pairings and pairings that are in contrast particularly high frequency, these measure choices and selection do not appear to be well explained in light of the broader picture that we obtain from examining the computational literature that has already explored and presented many other alternative measures (such as the developed narrative in Section 4.4).

Similarly, researchers with an interest in applying association measures to learner writing are left to wonder how the choice of the MI and t-score particularly 'fit' into this wider association measure picture and how these measures may be different to others mentioned in the literature. As already noted in Section 4.4, a further consideration here in this work is that few holistic narratives of association measures exist, and when in existence it proves a challenge for the reader to visualise how family member measures are strongly or weakly related, and how measures across families are in fact related. These issues mean measure selection for researchers is a fragmented affair with researchers seemingly following the path of others without exploring viable options from the informing computational literature. Although we can see a change in this with the introduction of the Delta P and Collexeme Strength, Gries and Durrant (2020) continue to note the lack of exploration in the field as a whole and how the fragmented narrative has thus far failed to encourage such exploration.

To this end, in an attempt to engage with this literature gap and potentially choose an association measure set that may be able to highlight different properties of collocation, the present study has the aim of visualising the similarities and differences between these measures of association by way of a

cluster analysis. This engagement with the literature gap is necessary to not take a position that simply relies on the MI and t-score but to clarify how possible measures of association are related and distinct so as to choose measures that offer the opportunity to tap into different properties of collocation. Therefore, the cluster analysis becomes a way to establish connections and intricacies between the potential measure set and make better informed decisions on measure selection from this information.

The review now turns to consider two other valid gaps in the literature that need to be taken into account in any measurement of the relationship between collocation and writing quality.

4.7 Course and Learner Variables in Feature-Quality Relationships

The relationship between language features and writing quality appears to be straightforward. However, Sudweek, Reeve, and Bradshaw (2005, p. 240) point out how a student's essay score is shaped by several factors including the task, rater variables and the student's background. Barkaoui (2008) also emphasises that the awarding of grades is influenced by a plethora of rater and contextual assessment factors.

These factors mean that the awarding of grades is a complex practice that is steeped in variation: a practice that is not as simple as claiming that writers are awarded higher or lower grades simply as a function of their use of linguistic features, or indeed collocations. As Barkaoui (2008) helps indicate, the assessment community has traditionally held two views on this variation: one that sees this variation as a source of error (See Gere, 1980; Charney, 1984; Huot, 1990a, 1990b for discussion of this position) and one that sees this variation as worth exploring as the grading process remains a partly standardised, and

constrained endeavour (see Eckes, 2008; Deville & Chalhoub-Deville (2006) for discussion of this position).

In examining the relationship between collocations and writing quality, there is a need to recognise that there are a number of other variables that have an equal theoretical chance of contributing to writing quality judgements. A number of studies have highlighted how task and topic are two such variables (e.g., Bouwer, Beguin, Sanders and van den Beurgh, 2015; Hake, 1986; Quellmalz, Capell & Chov, 1982; Ruth & Murphy, 1988; Spaan, 1993; Tedick, 1990) as well as the language status of the writer being an equally plausible source of variation (e.g., Brown, 1991) The sub-sections that follow bring together this evidence and highlight how it introduces variability into the relationship between collocations and grades. Section 4.7.1 examines the role of task and topic while Section 4.7.2 looks at the role of writers' language status in the influence of raters awarding grades.

4.7.1 Task and Topic Effects

Several studies across L1 and L2 contexts have noted how task differences can explain variations in grade scores (e.g., Bouwer, Beguin, Sanders and van den Beurgh, 2015; Hake, 1986; Quellmalz, Capell & Chov, 1982; Ruth & Murphy, 1988; Spaan, 1993; Tedick, 1990). The present study acknowledges this by drawing on a number of past and present seminal studies into this relationship. In a study of an L1 school context in the US, Quellzmalz et al (1982) found that across 200 11th and 12th grade writers' texts, raters awarded markedly lower scores to narrative texts over expository texts. The rationale for this variation is speculated to be that raters have a tendency to score narratives more harshly. Their study shows that level of performance varies on tasks casting doubt on the

assumption that a good writer is a good writer irrespective of assignment. In L2 contexts similar comments about task have also been made. Carlson and Bridgeman (1986, p. 141) note that our understanding of writing quality does not remain stagnant as tasks change. Hake (1986) is another who has questioned if different types of writing are graded differently.

Ruth and Murphy (1988) also discuss a number of task and topic effects on rater variation. They note how task and topic effects are underexplored in the research of rater and/or score variation. Carlson and Bridgeman (1986) broadly highlight the influence of topic while with L2 graduate students, Tedick (1990) found that over her sample of 105 graduate students, writing performance was higher on the field specific topic rather than the generic topic. Hamp-Lyons and Mathias (1994) is another similar study that looked at expert judgements and the assumptions that expert judges make about task difficulty and scoring. Importantly, Hamp-Lyons and Mathias (1994) found that experts' judgements did not always match scoring patterns and conclude that task effects can provide an important glimpse into rater variation. More recently, Johnson, Penny and Gordon (2009) also note the importance of task differences in influencing rater variation. Lavallee and McDonough (2015) also found with their EAP students, their effect essay prompt was scored lower than their cause prompts.

In a series of studies that focus on how raters may look for different linguistic features in awarding high grades, Guo et al., (2013), Kyle and Crossley (2016) have studied how lexical features across two TOEFL tasks have different correlations with writing quality. Guo et al (2013) studied both the independent essay and the integrated source-based essays from TOEFL and found that a number of lexical diversity and sophistication measures yielded different positive or negative correlations with the writing quality grade scores across the two essay

tasks. A similar finding was found in Kyle and Crossley's (2016) study of independent and integrated source-based TOEFL essays.

Taken together these studies provide evidence that there is the possibility that alongside the linguistic features in an essay, there is the possibility that certain tasks and topics may influence rater variation and introduce sources of variation and/or bias into the scoring process. These are issues that need to be considered in the FYC context since it includes more than one task and students are allowed to choose their own topics across tasks.

4.7.2 The Language Status of Writers

Other studies have accounted for the role that the writers' language background may have on essay score variability (e.g., Brown, 1991; Santos, 1988; Song & Caruso, 1996; Huang & Foote, 2010). These studies present a mixed picture that shows how language background influences the allocation of writing grades by raters. In a seminal study in the US, Brown (1991) studied a similar context to the FYC context in the present study. He looked at the degree of difference existing in the writing scores of 56 native and 56 international students studying composition courses at a US university. Results found no statistically significant differences between the two writers' groups scores; however, faculty did pay attention to different features of the writing, showing that although no major score differences existed, raters may have arrived at their scores from different perspectives. However, the graduate level study from Huang and Foote (2010) more recently found that L2 writers received consistently lower essay scores than their native counterparts, leading Huang and Foote (2010) to share concerns about score reliability for the L2 writer group. The studies show a mixed picture of how when rating essays raters may have conscious or unconscious patterns

of grading depending on whether or not the essays are written by first or second language writers. However, it is worth bearing in mind the largely qualitative way that this evidence has been found and the lack of generalisation that we can assume from these studies.

The inclusion of the variables of task, topic and writers' language status into a study of feature-grade relationships therefore is warranted to show how these underlying additional sources of variation may also contribute to grade scores. The review now turns to consider the final research gap of valid measurement via the use of appropriate statistical methods that can model the relationship between collocation and writing quality as well as the other valid variables mentioned in this section.

4.8 The Statistical Methods used to Capture Relationships

The previous section recognises that efforts have been made to acknowledge that the relationship between linguistic features and writing quality is not a simple one, and is instead influenced by task, topic and rater/writer influences. A further consideration in building up this picture of being able to tap into score variation is the structure of the corpus itself and how the sampling from the assessment context influences the variance in grade scores.

In this respect, when we examine previous feature-grade relationship studies, we see how these monofactorial studies are set up in such a manner that does not *fully* incorporate or acknowledge that raters are not static in their grading procedure. In the case of most FYC contexts, the volume of students in classes and enrolled on the whole programme means a single rater rates multiple essays and these essays may come from different classes. This grading situation means that the data points that make up the collocation-grade relationship are not

independent of each other because the fact that a single rater is responsible for a cluster of grades across more than one class means there is a dependency between the data points. Many previous feature-grade studies (e.g., Bestgen & Granger, 2014; Kyle & Crossley, 2016; Guo et al., 2013; Granger & Bestgen, 2014; Paquot, 2018; Yoon, 2016), have assumed that the data points are completely independent from all other data points included in the equation. In using multiple linear regression, scholars interested in measuring writing quality have made the assumption that the corpus of texts and related contextual and learner variables are independent data points or independent observations (Barkaoui, 2010). This can be seen in several studies that use large scale international exam scripts as their corpus. In the case of IELTS and TOEFL based corpora, the scripts used are graded by multiple raters and for analysis purposes their observed grades cannot be considered independent observations as the texts are clustered into different raters (e.g., Biber et al., 2016; Guo et al., 2013; Kyle & Crossley, 2018). The texts in a single cluster share similar characteristics e.g., being awarded a grade by the same rater. Several education methodologists have noted the implications of violating this assumption (e.g., Heck & Thomas, 2000; Hox, 2002; Osbourne, 2010). When these assumptions are violated, there is a greater chance of the Type I error rates being inflated (McCoach, 2010). At the same time, the individual writer is also a source of random variation because some students contribute more than one text to the corpus. This dependency also needs to be factored into the equation.

This violation of independence has implications for how we understand relationships between feature-grades in FYC programmes like the one at USF. The unique assessment structure means that if modelled with monofactorial

methods, we risk remaining unaware of the role that individual raters and writers play in modelling collocation - grade relationships.

4.8.1 Research Gap II: Recognising Appropriate Statistical Methods

The literature set out in Section 4.7 has identified two pertinent research gaps that past research in feature-grade relationships has not always accounted for. First, the inclusion of other sources of variability such as task, topic and rater biases and second, the need to account for the fact that the corpus where texts are taken from contain levels of dependency because grade allocation is shared across multiple raters who rate essays across multiple classrooms and the texts themselves are produced by more than one student. In this respect, the literature review highlights that these two gaps need to be addressed in the examination of the FYC context in this present study.

4.9 Summary

This review has unpacked the notion of collocation within a wider understanding of the term phraseology. The review has also underlined the importance of collocation to academic writing quality; and in doing so has set out how collocation has been operationalised in terms of diversity and sophistication. The first issue that the review pointed to was the validation of these measures in that thus far, the emerging studies have used only a narrow set of possible measures that have been introduced in other association measure driven studies. Given this contrast, the present study aims to explore this fuller measure set language; and more clearly focus on uncovering how these association measures are related and the extent they appear to capture different collocation properties or features.

The review concluded by outlining how this seemingly simplistic relationship between collocations and writing quality is influenced by variation

introduced by contextual and learner variables that play a role in the learning and assessment make-up of the FYC programme. The study also accounts for the variation introduced by these variables. The remainder of the study proceeds as follows: Chapter Five sets out the overall methodology of the study, Chapter Six carries out a cluster analysis of association measures with the aim of yielding a measure set which taps into different properties of collocation with these measures then informing the study of the relationships between collocations and writing quality. Chapter Seven carries out mixed-effects modelling to determine relationships between collocation and writing quality inclusive of both fixed effect task and writers' language status variables and random effects relating to individual raters and students. Chapter Eight concludes the study by outlining a number of implications for the measurement community and outlining how these results may move forward the FYC narrative to consider how future language instruction may be shaped on the programme.

Chapter Five: Overview of the Methodology

5.1 Introduction

This chapter details the two-study approach taken in the research. Overall, the study adopts a corpus-based research design that is based on examining the relationship between measures of collocation use and writing quality grades. The first study looks at collocation measure selection and at the relationships between these measures via a cluster analysis; and the second follow up study examines how those measures have a relationship with writing grades and how this relationship varies when learner and contextual variables are considered in mixed-effects regression modelling. The chapter begins by covering: the research design that is adopted, the project and text selection, generating a working corpus from the respective sampling frame, cleaning the corpus texts, the text pre-processing workflow that included parsing the texts to facilitate extracting collocation types and establishing the accuracy of the parser.

5.2 The Approach Adopted to Research Inquiry

Before setting out the research design, it is necessary to consider the underlying theoretical paradigm that guides the research. A paradigm is defined as a way of looking at the world, including a view of how research and science should be conducted (Creswell, 2014; Punch & Oancea, 2014). The research is influenced by the view that research inquiry and more broadly language use can be considered through a post-positivist lens. However, the approach adopted here is not solely contained by the paradigmatic stance; it also acknowledges that the research questions that are being asked play a crucial role in the approach taken to the research inquiry. Under Pring's (2014) views, research inquiry can be question-driven in the sense that the researcher asks probing questions that are pertinent to their field at that time and equally under this view, knowledge gained from this inquiry holds true at that time until it is built upon in further inquiries (Crotty, 2015; Pring, 2014; Howell, 2013). Post-positivism emanates from an ontological and epistemological vantage point that rejects the rigid tenets of positivism. Ontology concerns how reality is constructed and how it exists whereas epistemology, in contrast, focuses on how knowledge is viewed and constructed (Greener, 2011; Grix, 2004).

Positivism holds that there is an absolute objective reality where objects and subjects are separable and exist independently from each other (Howell, 2013). Post-positivists seek objectivity but do so by adopting a critical realist view of inquiry and an epistemology which appreciates that absolute truth does not exist (Rowbottom & Aiston, 2006). Post-positivists believe objectivity is only ever tentatively reached by acknowledging relevant contextual factors that contribute

to studying human behaviour (Creswell, 2014; Onwuegbuzie, Johnson & Collins, 2009). The research adopts a post-positivist stance in that it looks to discover how well measures of writing quality in the form of grade scores correlate to text-based measures of collocational features. It draws on learner and contextual factors by considering how the task type, language status and individual raters and classes influence this relationship in mixed-effect modelling.

The research focuses on the belief that quality writing has a relationship with a number of features that operate as shared commonalities between 'good' texts. This follows the belief of much previous work (e.g., Arthur, 1979; Crossley & McNamara, 2012; Ferris, 1994; Li, 1996; Perin & Lauterbach, 2018; Ruegg et al., 2011; Vann, 1979). In this research, there is an underlying assumption that collocation use can be indicative of writing quality and that these collocational features are shared across similar quality texts but also that certain collocational features distinguish between texts that are of different quality. This post-positivist approach is reductionist in nature meaning that in this research there is a recognition that the relationship between collocations and writing quality is reduced to a narrow set of key measures that continue to have some relationship to writing quality when contextual and learner variables are taken into consideration.

5.3 Research Design

The research comprises of a corpus-based design. The overall design is exploratory in nature because the research adopts an open position on the exact measures of collocation that may have a relationship to writing quality. The exploratory nature of the work is facilitated by two related studies.

The first study aims to show how the various measures of association are related and to what extent they may tap into similar collocation properties by way of their formulae. This aim is achieved by analysing measures of sophistication through a cluster analysis (thereby answering research question one). Using the results from the cluster analysis, the second study uses a traditional logistic ordinal regression model to determine the relationships between these association measures, a simple measure of diversity and the variables of task and language status and holistic grades (thereby answering research questions 2.1 and 2.2). Then a subsequent mixed-effects model is produced to look at how these relationships vary when the model accounts for the fact that individual raters are crossed with classes (thereby answering research question 2.3).

The remainder of this chapter is organised as follows. Section 5.4 outlines the data collection procedures, generating an appropriate corpus sampling frame and working corpus, text annotation and evaluating the chosen collocation extraction technique in relation to other more traditionally established extraction techniques. Section 5.5 details choosing an appropriate initial bank of diversity and sophistication measures and Section 5.6 outlines the extraction of dependency pairs and their allocation of association measures. The chapter concludes by outlining how these dependency pairings are then analysed in Chapters Six and Seven.

5.4 Data Collection Procedures

This section of the chapter clarifies the design of the FYC corpus including the treatment of texts prior to analysis.

5.4.1 Gathering Student Metadata and Ensuring Ethical Research

Ethical text access and treatment were adhered to throughout the research process. Prior to accessing the texts, the research plan went through IRB (Institutional Review Board) checks at the University of South Florida and was reviewed and approved by the First Year Composition IRB coordinator. The IRB letter of approval are detailed in Appendix D. Since the texts were already in circulation in the My Reviewers data warehouse, the University of Exeter's ethics board also confirmed that the IRB check was acceptable adherence to their own ethical research guidelines.

In addition to this IRB approval, it should be emphasised that USF students are invited to opt in or out of submitting their essays to the My Reviewers data warehouse. This consent procedure is set out on the My Reviewers platform and students are under no obligation to allow their essays to be included. The collected texts were therefore only written by students who opted in to allowing their texts to be used for FYC research. All included texts were anonymised by the My Reviewers team who removed student names, ID numbers, and instructor names from the essay cover sheet and the Microsoft Word document headers.

5.4.2 Project and Text Selection

The study focuses on two distinct projects[16]. The first project is ENC 1101's 'Project three: Joining the conversation' and the second project is ENC 1102's 'Project one: Finding common ground'. These two projects were chosen because they require students to create texts under different objectives. The ENC 1101 project required students to set out two opposing stakeholder views on their chosen topic while the ENC 1102 project required students to build on their knowledge from ENC 1101 and suggest, through extensive reading, a possible common ground or compromise that could be reached between the two groups of stakeholders (as outlined in Chapter Two Section 2.3.2.1). The inclusion of these two projects facilitates examining how task type influences the grading process with the decision to include these two projects grounded in support from the literature review set out in Chapter Four (Section 4.7) which indicates task type is likely to be a factor that influences grading and therefore needs to be considered when investigating the relationship between collocations and writing quality.

From these two projects, texts were sampled from Fall 2016 running from August 2016 – December 2016 and Spring 2017 running from January 2017 – May 2017. Text collection did not include the summer semester that ran through June and July 2017 because student enrolment was particularly low during this time. Texts were selected with the learner and contextual variables of language status, and task type in mind. In order to ensure that the texts sampled included these variables, My Reviewers metadata was consulted as described in the following section.

[16] The study uses the term project as it is consistently used in the FYC programme. Each project is equal to a task and so the terms task and project are used synonymously here.

5.4.3 Creating a Working Corpus from the My Reviewers Data Warehouse

This section of the chapter describes how a sample was taken from the data warehouse and how it was cleaned and prepared for the analysis in the cluster and analysis and modelling studies.

5.4.3.1 Using demographic data to establish a corpus sampling frame

An initial consideration in determining corpus size was how many writers had completed the demographic survey (as shown in Appendix E). The demographic survey yielded information about the language status of the writers and so the inclusion of learner variables was dependent on the successful completion of the survey. This had an important influence on the creation of the final corpus and the regression modelling process because missing data (e.g., from the non-completion of the survey) at one level of analysis could influence the inferences made at higher levels of analysis (McCoach, 2010) and potentially lead to convergence problems with the model (Bodo Winter, personal communication, 2018; Eager & Roy, 2017). Winter (2020,p. 267) encourages researchers to anticipate problems with convergence and missing data at the design stage of the study and so I made the decision, in the absence of clear-cut advice on how much missing data would be permissible without leading to measurement errors, I would model with a clean data set and include only complete cases where all data was present for the variables of interest. This limitation of inferencing should be kept in mind and will be revisited in Chapter Eight's conclusion to the study.

Using student responses to the demographic survey, language status was determined by comparing answers to the question: 'What is the first language that you learned?'. This question was explained to students as referring to their mother tongue (first language). In answering this question, students had several

options available to them including English, English and another language, and several language choices which indicated English was not their first language (e.g., Mandarin, Spanish, Arabic and Russian). In examining these answers, I had the goal of creating two distinct writer groups that differed in their English proficiency in a markedly different way. Based on these answers, texts were filtered to exclude texts written by students who identified as speaking 'English and another language' because this would have placed students in between the two exclusive groups of English as a first language (L1) and English as a second language (L2). This split mirrors the similar interest in this dichotomy that other FYC researchers have had. For example, Eckstein and Ferris (2018) who studied L1 and L2 differences across computational measures of syntax and lexis. However, unlike the unexplained division in this FYC example, the dichotomy in this research is explicitly and exclusively driven by the self-reported language status of the student writers**.** This meant that for ENC 1101 project 3, 709 final draft texts were extracted from My Reviewers for Fall 2016 because these writer texts had fully or partially complete demographic surveys. An examination of the metadata revealed that 256 writers did not answer the first language question while 72 writers answered that their first language was 'English and another'.

Similarly, for ENC 1102, 673 final draft texts were extracted from My Reviewers for Spring 2017. Another examination of the metadata revealed that more writers in ENC 1102 answered the demographic survey and therefore only 19 writers were excluded for non-completion while 93 possible bilingual writers were also excluded. Finally, grades D (D, D- and D+) and grade F texts were removed across ENC 1101 and ENC 1102 modules. This filtering was in keeping with the research aim of examining texts that achieved the minimum passing grade set out under Florida's state education requirements which explain that

enrolled students must achieve a minimum grade of C- to pass the FYC programme (Undergraduate Core Curriculum Information, 2018). The self-reporting of language status by the students themselves has both advantages and disadvantages. On the one hand, it is an efficient way of gaining an insight into how they perceive their language status. However, on the other hand, self-reporting language status or ability has raised validity concerns (e.g., Oscarson, 1989; Tomoschuk, Ferreira & Gollan, 2019) because the students' evaluations are often overly subjective and not guided by literature on language status. The reliance on self-reporting language status in this manner therefore should be kept in mind as a potential limitation of the research that follows.

This filtering produced a final corpus of 879 texts across the two module projects: 366 texts for ENC 1101 project 3 (labelled as 'Task 1') and 513 texts for ENC 1102 project 1 (labelled as 'Task 2'). The final corpus totalled 1,035,319 words (397,809 words in the ENC 1101 sub corpus and 637,510 words in the ENC 1102 sub corpus) and is described in Tables 17 - 19 :

Table 17: General Corpus Make-Up

Module	**Project**	**Number of texts**		**Words per L1 texts**			**Words per L2 texts**		
		L1	**L2**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
1101	3	262	104	1,082	615	2,279	1,100	655	1,849
1102	1	404	109	1,237	529	2,457	1,251	784	1,951

The general corpus make-up in Table 17 shows that from the demographic survey more students identified as L1 speakers of English with a smaller population identifying as L2 speakers. Tables 18 and 19 show the number of texts at each grade level (A-C) across L1 and L2 populations. Tables 18 and 19 show

most students achieved a grade of A+ to B+ with fewer students achieving C grades:

Table 18: ENC 1101 Grade Breakdown

ENC 1101				
Grades	**# L1 texts**	**# L2 texts**	**Total texts per grade**	**Percentage (%) of total texts**
A+	24	21	45	12.30
A	34	12	46	12.57
A-	62	26	88	24.04
B+	46	14	60	16.39
B	36	9	45	12.30
B-	27	11	38	10.38
C+	17	5	22	6.01
C	10	0	10	3.28
C-	8	4	12	2.73

Note: (# = number)

Table 19: ENC 1102 Grade Breakdown

ENC 1102				
Grades	**# L1 texts**	**# L2 texts**	**Total texts per grade**	**Percentage (%) of total texts**
A+	53	19	72	14.04
A	63	19	82	15.98
A-	84	14	98	19.10
B+	59	13	72	14.04
B	56	20	76	14.81
B-	32	12	44	8.58
C+	21	6	27	5.26
C	15	5	20	3.90
C-	20	2	22	4.29

Note: (# = number)

Tables 20-21 also which show the analytical and holistic grade breakdown indicate that the mean scores for each analytical grade (style, organisation, format, analysis and evidence) are also relatively high with students in L1 and L2 populations achieving a mean of ≥ 6 points (from a maximum of 8 points) for each component. Similarly, mean holistic grades for each population are above 11

(approaching a B+). This shows that overall scores indicate that many students achieved high passing grades rather than a simple pass.

Across the two tasks students were free to choose their own topic. However, in line with the learning outcomes of the FYC programme, this topic was expected to be developed across modules with students who completed both modules expected to retain the same topic throughout the programme. However, in some cases, some students only completed one module (ENC 1101 or ENC 1102) and so topic choices were not always uniform across the texts in the corpus. Student topic choices were wide ranging across both learner groups with few possible found across tasks or language groups. The two learner groups both had students who wrote about controversial international topics such as '*abortion*', '*social anxiety*' and '*renewable energy sources*'. While common topics across the L1 population focused on U.S domestic issues (e.g., '*immigration in the US*', '*legalising marijuana*' and '*the abuse of over the counter medication*'), some L2 international students wrote about their own domestic issues (e.g., '*poverty in Honduras*' and '*economic growth in China*'). In ENC 1101, students wrote essays that presented both sides of the arguments surrounding their topic with some students, but not all, showing a clear conclusion that supported one side of the argument. In ENC 1102 students set out to offer a 'common ground' and presented a more balanced look at their topics with the aim of suggesting how scholars on both sides could reach a compromise in their views.

Table 20: Grading Breakdown for L1 Students

Module	Grade breakdown											
	Holistic		Style		Organisation		Format		Analysis		Evidence	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ENC 1101 Project 3	11.97	2.13	6.39	2.06	6.38	2.20	6.21	2.19	6.22	2.17	6.02	2.27
ENC 1102 Project 1	12.07	2.24	6.89	1.29	7.17	1.21	6.95	1.40	6.87	1.29	6.72	1.44

Table 21: Grading Breakdown for L2 Students

Module	Grade breakdown											
	Holistic		Style		Organisation		Format		Analysis		Evidence	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ENC 1101 Project 3	12.69	2.02	6.37	2.42	6.42	2.49	6.30	2.53	6.31	2.49	6.13	2.60
ENC 1102 Project 1	12.01	2.24	6.82	1.17	7.22	1.14	7.09	1.26	7.02	1.17	6.99	1.27

5.4.3.2 Raw text conversion

Essays were downloaded by the My Reviewers development team[17] via the My Reviewers interface and stored in course batches in their in-house data warehouse. Since the My Reviewers platform deals with PDF files to allow students and instructors to use PDF mark-up tools when they offer review comments and allocate grades, essays were downloaded in PDF format and then converted to text files (.txt). This conversion to text files was necessary to allow easy pre-processing and collocation extraction to take place using custom written R scripts (R Core Development Team, 2014).

The text files went through several rounds of formatting to ensure reliable feature counts and accurate processing in R. In their raw state, the texts contained several features which were problematic to the goal of studying collocations. These problematic features included bibliographies or reference lists, which, if included in the analysis, would have deflated collocation frequency counts and provided a skewed picture of learners' collocation production because texts would be longer and therefore normalised frequency counts would be lower across more words. These lists were easily identified because of their predictable appearance at the end of essays and their precluding header: 'references', 'reference list', 'bibliography' or 'works cited'. These lists were automatically removed from texts by using an R script[18] which generated a 'working corpus' that contained texts with removed bibliographies and reference lists (Durrant et al., 2019)

[17] This team included faculty at USF including Dr. Joseph Moxley, Rajeev Reddy Rachamalla, Dat Be Le and Dhairya Dave.

[18] The R script was written by Dr. Philip Durrant as part of the co-authored cited publication.

Similarly, texts were also manually formatted to remove page numbers which appeared as numbers or included the word ‘page’ followed by a number. These were problematic as they appeared sporadically throughout the texts, because converting the texts from PDF to text files distorted the position of the document header, and this meant the header information appeared across clause, phrase and sentence boundaries. In example (1), the word ‘page’ and number ‘2’ were manually deleted.

(1) ‘Ivory today has a black market value of **page 2** $100,000 for every kilogram and is still rising [ENC 1102]’.

The document header also contained markers of anonymisation (denoted by a ‘_’ marker) that the My Reviewers team had used to anonymize the texts. The document header and these markers were also removed to allow the Stanford Core NLP tools (Manning et al., 2014) to operate more efficiently. Since the marker was not a recognised punctuation convention it may have confused the parser when trying to identify dependency relationships between words (Huang, Murakami, Alexopoulou & Korhonen, 2018).

5.4.3.3 Text pre-analysis workflow

This section explains the pre-analysis that texts went through before the two studies could be conducted. The steps are set out in Table 22 with the initial three steps of lemmatisation, POS tagging and parsing explained first as they occur together when using the Stanford Core NLP tools then the extraction of dependency pairs, the computation of frequencies and allocation of association measures follow.

Table 22: Text Pre-analysis Workflow

Steps	Tools	Corpora used	R script
Lemmatisation Part-of-speech (POS) tagging Parsing	Stanford Core NLP annotators	FYC corpus MICUSP	Parsing script (Script One)
Tidy up parsing format	R programming	FYC corpus MICUSP	Tidy up parse script (Script Two)
Extraction of dependencies	R programming	FYC corpus MICUSP	Extraction script (Script Three[19])
Manual parser check	Manual annotation	FYC corpus	No script used
Compute corpus-based frequencies	R programming	FYC corpus MICUSP	Frequencies script (Script Four)

To eventually calculate association measures for each dependency pairing, a number of preliminary steps were carried out on both the study corpus (the 'FYC corpus' hereafter) and the reference corpus MICUSP ('The Michigan Corpus of Upper Student Papers') (Michigan Corpus of Upper Student Papers, 2009). The reference corpus was used to compute a range of association measures by calculating observed and expected frequencies. MICUSP was selected as the reference corpora because, it could feasibly act as a reference for FYC students in the sense that the range of tasks and topics represented were similar to the types of tasks that these writers would be expected to complete in their academic studies after completing the FYC programme. These tasks included argumentative writing similar to some of the FYC module tasks (e.g., stakeholder analysis) but also included literature reviews, lab reports, critical summaries of texts and empirical research reports. The breakdown for genre type was decided by two raters, with 44% of texts being reports, 22% of texts being

[19] Scripts One to Three were written by Dr. Philip Durrant as part of the cited publication Durrant et al (2019).

argumentative writing, 17% being research papers, 7% being critiques, 6% being proposals, 3% being response papers and 1% being creative writing. The structure of the corpus and the range of tasks therefore seemed to represent a corpus that was indicative of the types of writing that the FYC programme aimed to prepare students for writing. The second reason for choosing MICUSP related to its size with the corpus itself representing one of the largest corpora of student writing across multiple levels of study (undergraduate and graduate), disciplinary subject and language backgrounds (in alignment with the study corpus, MICUSP also included non-native student writers (e.g., Arabic, Chinese and Spanish L1 speakers). The free access to circa 2.8 million words of student writing would therefore also allow confidence that the reference corpus was a good representative of authentic student writing which acts as a target for FYC writers.

However, a caveat worth mentioning here relates to control of ensuring papers were upper-level grades. In their summary of MICUSP, Römer and O'Donnell (2011) note that they largely depended on the student's assertion that their voluntary submission to the corpus had received an A- or A grade from their instructor. Although, Römer and O'Donnell request the name of the instructor, it is not documented how many texts were verified by asking the instructor for confirmation of grade and it's also not elaborated in their summary or on the MICUSP repository, how writing grades were decided, according to particular writing rubric measures. This is one drawback of using MICUSP.

The use of the corpus complied with the MICUSP Fair Use Statement and the formatted texts that were used to compute association measures were solely based on anonymised metadata[20]. I cleaned up the texts by removing any titles,

[20] The Fair Use Statement reads: "The Michigan Corpus of Upper-level Student Papers (MICUSP) is owned by the Regents of the University of Michigan (UM), who hold the copyright.

reference lists and references to figures and mathematical formulas to mirror similar formatting in the FYC corpus but also to preserve actual student produced language in the sense that texts may have included third-party tables, charts and graphs, formulas or any other material.

A summary of the make-up of the MICUSP corpus is presented in Tables 23 and 24:

The corpus has been developed by researchers at the UM English Language Institute. The corpus files are freely available for study, research and teaching. However, if any portion of this material is to be used for commercial purposes, such as for textbooks or tests, permission must be obtained in advance and a license fee may be required. For further information about copyright permissions, please contact Dr. Ute Römer" at elicorpora@umich.edu. Statement is available at: https://micusp.elicorpora.info/

Table 23: MICUSP Corpus Make-up

Academic Subject	Sample topics
Biology	• Fruit fly experiments
Economics	• The economic recession • Economics of the illicit drug-market
Education	• The No Child Left Behind Act • Standardised testing
Civil and Environmental Engineering	• The use of reinforced concrete shear walls in steel framed buildings. • International law and environmental policy
History	• New Social History • Sex education in East and West Germany
Industrial and Operational Engineering	• Developing a student transport plan in Downtown Detroit. • External analysis of the National Society of Black engineers.

Table 24: MICUSP Disciplines and Writer Backgrounds

Academic division	**Academic Discipline**	**# Texts (n=829)**	**# Words (2,367,652)**	**# L1 writers (n=681)**	**# L2 writers (n=148)**
Humanities and arts (n=223)	English (ENG)	98	253,580	90	8
	History and classical studies (HIS_CLS)	40	146,820	38	2
	Linguistics (LIN)	41	142,820	34	7
	Philosophy (PHI)	44	124,500	39	5
Social sciences (n=309)	Economics (ECO)	25	66,320	16	9
	Education (EDU)	46	137,502	42	4
	Political Science (POL)	62	196,543	53	9
	Psychology (PSY)	104	290,310	81	23
	Sociology (SOC)	72	205,049	51	21
Biological and health sciences (n=171)	Biology (BIO)	67	148,433	57	10
	Natural Resources (NRE)	62	154,348	57	5
	Nursing (NUR)	42	143,694	30	12
Physical sciences (n=126)	Civil and environmental engineering (CEE)	31	85,952	25	6
	Industrial and operations engineering (IOE)	42	119,271	27	15
	Mechanical engineering (MEC)	32	110,675	22	10
	Physics (PHY)	21	41,835	19	2

Note: The 2,367, 652 words are after text clean-up of the original texts.

Texts were lemmatised, POS (Part-of-Speech) tagged and parsed by running the Stanford Core NLP annotators through both the study and reference corpus. The Stanford tagger's default tagset originates from the Penn Treebank project (Santorini, 1990). The Stanford tag annotator tagged each text for each word's POS. This POS information was used as the starting point for the Stanford parser in deciding on the dependency relationships between the words on a sentence-by-sentence basis. This was carried out in command prompt using the following code referred to as Script One in Table 22:

```
java -cp "*" -Xmx10g edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators tokenize,ssplit,pos,lemma,depparse -fileList filelist.txt -outputDirectory output -outputFormat conllu
```

This code applied the annotators of tokenize, sentence split, part of speech tagging, lemmatisation and dependency parsing to all the files in the corpus listed in the "filelist.txt". The output was then saved in the same directory under a folder named 'output'. The output was saved in Conllu (.conllu) format (de Marneffe & Manning, 2016). The original output in .conllu format was tidied up to become more user-friendly and easily readable by using an R script (labelled as Script Two in Table 22) that tidied up the output and effectively grouped the information under the following column headings: "sentence number", "word number", "word", "lemma", "POS", "dep_on" and "dep" to show the sentence number, each word number in the sentence, its lemma form, part of speech, dependency relation by using the word number ("dep_on") and its dependency tag ("dep"). This output is shown in Table 25

Table 25: Sample Parser Output

Sentence ID	Word_number	Word	Lemma	POS	dep_on	dep
1	1	Wearing	wear	VBG	7	csubj
1	2	good	good	JJ	3	amod
1	3	clothes	clothes	NNS	1	dobj
1	4	is	be	VBZ	7	cop
1	5	the	the	DT	7	det
1	6	best	best	JJS	7	amod
1	7	way	way	NN	0	root
1	8	to	to	TO	9	mark
1	9	show	show	VB	7	acl
1	10	charisma	charisma	NN	9	dobj
1	11	and	and	CC	9	cc
1	12	leave	leave	VB	9	conj
1	13	a	a	DT	15	det
1	14	good	good	JJ	15	amod
1	15	impression	impression	NN	12	dobj
1	16	to	to	TO	17	case
1	17	others	other	NNS	12	nmod
1	18	.	.	.	7	punct

Within this view of grammar, there are a number of classifications that linguists may question depending on their theoretical orientation. This includes the broad application of adjective 'JJ' and noun 'NN' tags. These broad applications can be seen in the tagging of semi-determiners including "other" which is repeatedly tagged as an adjective 'JJ' while indefinite pronouns like "someone" are classified as regular 'NN' nouns. Another point of contention that linguists may question relates to the tagging of the pronoun "one" which was repeatedly tagged as a cardinal number. Examples [2] and [3] illustrate these observations across different L1 and L2 texts:

[2] '*One* should carefully think before they eat fast food'

[3] 'There are three things that *one* needs to survive; food, shelter, and water, every living organism on the planet seeks food to get through life'.

These contrast with examples [4] and [5] which show instances when the tagging is appropriate across L1 and L2 texts:

[4] Poverty, in its most simplest form, can be defined as a situation in which *one* can not adequately meet and support his or her own basic needs with the resources available to them.

[5] When *one* thinks about cars, it is an immediate thought to think about motors, gasoline, and pricey maintenance.

In examples [4] and [5], 'one' is correctly identified as 'PRP' as it is a personal pronoun. However, further tag inaccuracies arise in the case of distinguishing between adjectives, adverbs, nouns and verbs. Sentences [6] – [8] show inaccurate tagging with sentence [6] illustrating how the original 'JJ' tag for

'individual' should be 'NN'; sentence [7] illustrating how the original 'NNS' tag for 'states' should be a verb ('VBZ': third person singular present); and sentence [8] illustrating how the original 'NN' tag for 'addresses' should be a verb ('VBZ'):

> [6] 'Adolescents may look at an *individual* and wonder why it is they themselves cannot have the life the individuals on social media have'.
>
> [7] 'The 14th Amendment in the United States constitution *states* that anybody who is born within the borders of the United States, has the right to become an American citizen'.
>
> [8] 'An article called the Role of Communication Technology in Adolescents Relationships and Identity Development *addresses* that on Facebook triggers states of envy and resentment in many'.

Across the categories of dependencies that the thesis focused on, there were a number of mistagged combinations that came to light when closely examined. These are highlighted in examples [9] – [12].

[9] He too **committed suicide** but this time it was decades after he received his last concussion (where committed is tagged as an adjective and the combination marked as an amod dependency).

[10] A study at the Centers of Disease Control and Prevention estimated that day 160,000 kids **nationwide stay** home from school because afraid of being (Karlinsky, Wright IV) ('stay' is tagged as a noun when it should be a verb).

[11] **Switching** to vegetarianism can **save** not only the environment but also can help improve health. ('Switching is tagged as a NN and so 'switching save' is tagged as a nsubj

dependency).

[12] NASA has performed numerous studies and have taken **photo evidence** to prove that the planet is indeed changing as a consequence to human activity. ('Photo' is tagged as an adjective here, so 'photo evidence' is treated as an amod dependency).

In generating parsed output, the Stanford parser provides lemmatised output as its default position. The decision to perform lemmatisation is not without criticism as several corpus linguists have highlighted that lemmatisation masks important collocates that inflected word forms have as unique pairings (Hoey, 2005) Hoey (2005) sums up some of the key scholarly arguments from Sinclair (1991) and Tognini-Bonelli (2001) in that they argue that collocational analysis should be performed on individual words because each word can have unique or 'special' collocational behaviour, for example, Hoey (2005), draws on the biology examples from Williams (1998) who notes that "gene" has different collocational behaviour dependent on its form . However, in this research, data analysis is carried out at the level of lemmas because as Evert (2004) and Seretan (2011) have pointed out, there is a risk of increasing error rates when calculating association measure scores, if these scores are based on pairings which occur too infrequently. Therefore, for this reason, the analysis is carried out on collocation types with these types appearing in the reference corpus a minimum of 5 times, as has been established as standard practice in other association measure studies (e.g., Bestgen & Granger, 2014; Chen, 2019; Granger & Bestgen, 2014; Paquot, 2018, 2019; Yoon, 2016).

5.4.4 Extracting Dependencies and Checking Parser Accuracy

5.4.4.1 Obtaining the dependencies from the parser output

After the texts had been passed through the parser, the next step was to extract the dependency types of interest. For this extraction, I used the custom-written script described in Durrant and Brenchley (in press), labelled as Script Three in Table 22. In brief, this script was designed to read through each text and extract amod, advmod, nsubj and idobj/dobj dependencies through a function labelled 'dep.pairs'. The 'dep.pair' function operated as follows. For amod dependencies, the function searched for rows that contained 'JJ' (adjective) in the POS column, a 'amod' in the dep column, a 'dep_on' value that attaches or points to a word with 'NN' (common noun) in the POS column. The function then records the dependent adjective and the noun it depends on. The script went through the same search process for advmod (both 'JJ', 'RB' : Adverbs with a modifying dependency on an adjective; and 'VB' 'RB': 'Adverbs with a modifying dependency on a verb'), nsubj ('VB', 'NN': 'Common nouns with a subject (nsubj) dependency on a verb') and iobj/dobj ('VB','NN': 'Common nouns with an object (iobj/dobj) dependency on a verb') dependencies to provide records of all of the extracted dependencies for these syntactic relations.

5.4.4.2 Checking parser accuracy

Although the general landscape of feature quality studies has relied on extracting linguistic features via automatic taggers and parsers that can establish part of speech in the former and syntactic dependency relations in the latter, Meurers and Dickinson (2017) point out the challenges in using these tools with L2 data and Huang et al (2018) highlight that only a few studies have determined the accuracy of these methods (e.g., Geertzen et al., 2013; Van Rooy & Schafer,

2003,2009) with studies now beginning to carrying out initial validation checks before using these tools (e.g., Durrant & Brenchley, in press).

Across the recent literature that has relied on taggers and/or parsers, some have used automated tools that have taggers/parsers embedded in them. For example, the suite of tools from Kyle and Crossley (2019) which have taggers and parsers (the Charniak (Charniak, 2000) and Stanford parsers (Manning et al., 2016)) embedded into them have been used to study aspects of lexical sophistication (e.g., Guo et al., 2013), cohesion (e.g., Green, 2012) and syntax (e.g., Kyle & Crossley, 2018), while others have simply used the tagger (e.g., Granger & Bestgen, 2014) or parser (e.g., Paquot, 2018,2019) directly.

In studying collocations, the accuracy of taggers has a more established history in feature-quality studies however many of these studies have simply referred to their general accuracy established on unrelated datasets rather than carrying out checks on the dataset used in their own studies (e.g., Granger & Bestgen, 2014 take this approach). Parser accuracy has a more limited history with learner language as pointed out by Huang et al (2018). Instead of checking this accuracy, it is simply assumed that using the parser allows researchers to more accurately study lexical collocations whose words have a dependency relationship as opposed to studying word pairings identified by the tagger where the tagged words may simply appear in close proximity to each other without sharing a dependency relationship thus being guided by Evert (2004) and Seretan's (2011) observations. However, this approach can only be claimed to be reliable if the accuracy is checked and in failing to do so, the observations of extracting 'syntactic noise' with simple tagging raise the possibility of potentially extracting word combinations that have no dependency relation.

This lack of confirmation leads to questions of validation because the accuracy of the parser influences claims of measurement or construct validity including how accurate the parser is at determining the dependency relationships that it claims to be able to detect (Huang et al., 2018).

The Stanford parser was trained on data from the Wall Street Journal which contains native speaker texts and is limited to a specific genre therefore the use of the parser on often inaccurate learner or student English needs to be determined (Huang et al., 2018). This is especially important for the study's aims of measuring collocation features. A failure to record parser accuracy influences the extent the findings are valid and reproduceable as we cannot be reassured the features and their measures are actually based on extracted lexical collocations whose words share a dependency relationship. With these issues in mind, an initial part of carrying out steps 1-4 in Table 22 involved checking the parser accuracy before the automatic extractions from the R scripts were used to calculate the association measures.

In undertaking this accuracy check, I asked a second annotator to check the POS tagging and dependency accuracy. The second annotator held an MA in TESOL and Applied Linguistics and had worked at a university in China where she taught courses in English grammar teaching. She signed an annotator agreement form (See Appendix F) to ensure she understood that the texts could not be shared externally and that she was comfortable with the annotation task. Prior to annotation, we met to discuss the task and how the accuracy should be recorded. Prior to the second annotator coding the texts, I independently coded 10 texts per grade level which amounted to at least 10% of the total corpus sample (N=879 texts). I first checked the accuracy of the POS tags with the rationale being that these tags were then used to determine dependency relations

by the Stanford parser annotator. I then manually recorded all dependency pairings that matched the dependency types under study. The threshold of 10 texts per grade level was decided so as to be compatible with the few dependency checks of learner English already in existence (e.g., Huang et al., 2018). The second annotator then coded the 10% of the same texts (amounting to 18 texts) for POS tags and dependencies and we compared our coding. In checking the POS tags, we reached a fairly high level of agreement across POS tags (92%) but our agreement for the parser dependency check was less (86%). Since my coding was consistent enough with the second annotator, I then went on to code the remaining 162 texts independently. I compared my coding of the dependencies with those found by the parser. These results are presented in Table 26.

Across the five dependency types, Table 26 shows the performance of the parser across grade level, rounded to the nearest decimal place. Precision scores were calculated by dividing the number of 'true positives' by the number of 'true positives' + 'false positives' while recall scores were calculated by dividing the number of 'true positives' by the number of 'true positives' + 'false negatives'. False positives relate to the number of instances where a dependency is labelled as one type but it should be another type of dependency (e.g., misclassifications of some nouns and verbs) whereas false negatives related to dependencies that were not labelled 'X' but should have been (again typical examples include noun and verb misclassifications). Table 26 helps highlight how some of the discrepancies that are highlighted by this manual checking arise from adverb modifying syntactic types. The relatively low recall scores mean the parser was unable to capture many true adverb dependencies and so given concerns about the types of final inferences we could draw based on these frequencies; this

dependency type was not considered further in the research. These findings corroborate similar checks from Durrant and Brenchley (2020). In their parser check on texts written by school children in England, Durrant and Brenchley (2020) found parser precision to average at 80% and recall to average at 85% for amod dependencies, while for dobj dependencies, precision averaged at 76% and recall averaged at 24%. An emerging theme from Table 26 is that the scores for L2 texts are consistently slightly lower than those found for L1 texts. This is an interesting observation that is worth following up on in further research when looking into the performance of the parser. In this light, further research into this observation may follow along the lines of Huang et al (2018) who studied the relationship between learner errors and parser and tagging errors to determine where were learner errors related directly to errors made by the parser. This line of research could be an emerging justification for the patterns found here in the FYC texts.

Table 26: Precision and Recall Percentages (n=180 texts)

		Amod			Adv mod with adjective or verb			Nouns with subject dependency on a verb			Nouns with an object dependency on a verb (dobj/iobj)		
Module	**Texts**	**N**	**Recall**	**Precision**	**N**	**Recall**	**Precision**	**N**	**Recall**	**Precision**	**N**	**Recall**	**Precision**
ENC	L1	1,654	88%	83%	897	22%	79%	1,267	88%	83%	608	54%	82%
1101	L2	1,501	83%	77%	711	18%	75%	1,157	85%	82%	493	48%	76%
ENC	L1	1,847	85%	80%	945	24%	77%	1,562	86%	84%	841	42%	88%
1102	L2	1,830	81%	76%	890	19%	76%	1,471	80%	82%	803	36%	77%
Average across all texts		**1,708**	**84%**	**79%**	**861**	**21%**	**77%**	**1,364**	**85%**	**83%**	**686**	**45%**	**81%**

The parser check suggests a number of interesting observations and challenges when we revert back to the claims made by Seretan (2011) that parsing can filter out syntactic noise and better target real or true dependencies. When we consider the relatively low recall scores for adv_adj or adv_verb modifying dependencies, we can see how these claims do not always hold true.

Although based on a total sample of 180 texts across the FYC corpus, for the other syntactic types, precision and recall figures are still at an acceptable level to instil confidence in the Stanford Core NLP tools being used to generate and retrieve dependency pairings. In the case of object dependencies, although their recall scores are lower than the other dependency types, their inclusion still reaches an overall acceptable level with the caveat that the likely real number of these dependencies is likely to be much greater than what the parser is able to retrieve. In the case of adverb dependencies, the justification for exclusion is based on a much smaller number of retrievals as well as having issues of accuracies when they are in fact retrieved.

These dependencies from the Stanford parser were then analysed for their diversity and sophistication across texts. The rationale for the selection of diversity and sophistication measures is discussed in the next section.

5.5 Tapping into Collocation Complexity: Initial Measure Selection

This section of the chapter sets out the rationale for the diversity and sophistication measures that the study's analysis is based on. Section 5.5.1 discusses the rationale for choosing the Guiraud's index (Root TTR) as the measure of diversity. Section 5.5.2 discusses the rationale for selecting a subset of association measures that were introduced in Chapter Four (Section 4.4).

5.5.1 Diversity Measures

As explained in Chapter Four (Section 4.6.1), many measures are in existence for measuring diversity. However, in keeping with the few studies that have used dependencies as their method of extraction, I opted to use the same measure of diversity as they have used. Drawing on Paquot (2019), this meant I opted to use the Guiraud index (Root TTR) which measures diversity with the formula: T/√N. This calculation is the number of types 'T' divided by the square root of the number of tokens '/√N'. A second reason for this measure choice was that it was one of the few measures appearing in the few single word studies from FYC researchers (See Chapter Three, Section 3.4.1.2).

5.5.2 Sophistication Measures: Rationale for Inclusion in Cluster Analysis

Drawing on the literature in Chapter Four (Section 4.4), the rationale for choosing association measures was grounded in examining measures across the groupings identified in key work by Evert (2004), Pecina (2005, 2010) and Seretan (2011). The rationale was also guided by what had already been acknowledged in the collocation literature from L2 scholars who had studied the underlying properties of a few association measures (e.g., Gablasova et al., 2017; Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Bestgen & Granger, 2014; Paquot, 2018, 2019). At the same time, since the research would deal with

multiple association measures and their calculation, I also opted for choosing association measures which could be easily calculated without much ‘cost’ as Evert (2004) helps highlight. Under these guiding principles, I therefore decided to focus on measures that were labelled as hypothesis measures, mutual information-derived measures and coefficient measures. Within these groups, I excluded measures such as the Fisher-Yates as well as those measures labelled context measures in Pecina (2005, 2010) as these were computationally costly measures which would not be easily calculated across the many dependencies that were extracted. This decision is not to say that these measures do not hold promise for learning about the connections between association measures and collocation properties; the decision to exclude them is simply a practical one. This meant that the association measures in Table 27 were chosen as the starting point to carry out the cluster analysis.

Table 27: Selected Association Measures for the Cluster Analysis

Number	Measure	Number	Measure
1	Poisson-Stirling Log Measure	26	Yulle's *Q*
2	T-Score	27	Driver-Kroeber
3	Z-Score	28	Fifth Sokal Sneath
4	Chi-Squared Test	29	Pearson
5	Log Likelihood Ratio	30	Baroni-Urbani
6	Squared Log Likelihood Ratio	31	Braun-Blanquet
7	MI	32	Simpson
8	Relative Risk	33	Michael
9	Dice Coefficient	34	Mountford
10	Mutual Expectation	35	Fager
11	Jaccard	36	Unigram Subtuples
12	Geometric Mean	37	U cost
13	Minimum Sensitivity	38	S cost
14	Odds Ratio	39	R cost
15	Odds Ratio Disc	40	T Combined Cost
16	Russel-Rao	41	Normalised Expectation
17	Sokal-Michner	42	MI^3
18	Rogers-Tanimoto	43	Log Frequency Based Mutual Dependency
19	Hamann	44	Mutual Dependency
20	Third-Sokal Sneath	45	Salience
21	First Kulczynsky	46	Delta P W2 | W1
22	Second Sokal Sneath	47	Delta P W1 | W2
23	Second Kulczynsky		
24	Fourth Sokal Sneath		
25	Yulle's ω		

5.6 Measure Calculation

This section describes the calculation of the measure of diversity and the measures of sophistication in the form of the association measures.

5.6.1 Calculating the Root TTR

The Root TTR was calculated by using a simple R script which used the frequency information gathered from the earlier extraction and tallying up of extracted dependencies. The first script calculated the number of types and tokens for each dependency per text and then the second added on script used this information to calculate the Root TTR for each individual text.

5.6.2 Extracting Pairs from the MICUSP Corpus

Using the same basic script that extracted the dependencies in the FYC corpus (See Chapter Five, Section 5.4), I obtained the dependencies and frequency information (types and tokens) for the MICUSP corpus. This information was also obtained at the unit of each individual text, with the information for each text then totalled to give a sum of types and tokens for the complete corpus. This meant that for each dependency type, I obtained the following information:

- The number of types and tokens for each dependency in each text and for the whole MICUSP corpus.
- The frequency of each individual word in each dependency type.

This information was eventually used to calculate the association measures. A point to reiterate here is that the dependency analysis for the cluster analysis and the final mixed-effects model was based on the extracted dependency type and tokens as lemmatised counts. This was done to ensure that association measure calculations were based on robust enough frequency counts (e.g., see the argument made in Evert, (2004) on the implications of basing analysis on infrequent, unstable or rare frequency counts).

5.6.3 Absent and Below Threshold Combinations

Using the frequency information described in Section 5.6.2, I then set out to exclude absent and below threshold combinations from the analysis. Absent combinations were those in the FYC corpus but not found in the MICUSP corpus; and below threshold combinations were those found in both corpora but they occurred in the MICUSP corpus fewer than five times. This exclusion was carried out in R by matching up a list of the FYC combinations and a list for the MICUSP combinations. It is important to recognise that following this threshold approach from past literature (e.g., Paquot, 2018, 2019) means that the picture obtained of learners' use of combinations is constrained by what is found in native corpora. The implications of this decision are that the analysis presents a picture of use, that does not include low frequency, idiosyncratic uses, but instead focuses on combinations found in the native reference corpus. This is an important limitation to acknowledge here.

A snapshot of below threshold and absent dependencies are shown in examples 13 (below threshold combinations) and 14 (absent combinations):

[13] universal__jj_:_ healthcare__nn,

cause__vb_:_ people__nn,

teacher__nn_:_ do__vb,

[14] athlete__nn_:_ receive__vb,

absurd__jj_:_ essay__nn,

giant__jj_:_ panda__nn,

Many of these combinations related to the task influences and differences between the two corpora. A more detailed picture of these combinations is set out in Chapter Six (Section 6.2) when the entries in the cluster analysis are described.

5.6.4 Calculating Association Measures

While Paquot (2019) specifically used in house custom written Perl scripts to calculate her association measures, I decided to opt for more manual control of the calculation. This decision was led by the fact that the research involved multiple association measures and multiple dependency types and I wanted to ensure that this calculation remained accurate over such a vast amount of dependencies and association measures. At this early stage of the research, I also wanted to manually make observations about the values being calculated for different dependency types, with the thought process being that actually inspecting this process may help highlight interesting patterns which I may later go back and discuss in the final analysis. Therefore, I chose to calculate the association measures by using the association measure calculator downloadable from Durrant (2020)[21]. This allowed me to set out all of the information needed to

[21] The calculator was retrieved from: https://phildurrant.net/resources/. Last accessed: 27.07.2020.

calculate association measures for the 47 measures efficiently. The exact calculation of the measures is described in Chapter Six when describing the steps in the cluster analysis.

5.7 Summary

This chapter has outlined the general methodology that the thesis adopts. The chapter has helped highlight the corpus design, sampling procedures and initial text treatment prior to conducting the cluster analysis and multi-level modelling. The next chapter sets out the procedure for conducting the cluster analysis, its respective results and what these results mean for modelling multi-level interactions between measures of collocation sophistication, writing quality and specific learner and contextual variables.

Chapter Six: Cluster Analysis

6.1 Introduction

As indicated in Chapter Four (Section 4.4), the association measure literature which informs measure selection in first and second language is often fragmented with language researchers in both camps drawing on a small number of past studies to frame their measure selection around. Embedded within this literature is a narrow range of the possible association measures that researchers may use in their exploratory/explanatory work. Given this rather limited view of association measures, this study has the initial aim of essentially taking a step back and re-evaluating how the association measures that are scattered around the literature may be similar and/or distinct and examine how much the mathematical properties of these measures may actually illuminate different types of collocation properties. In achieving this aim, it is hoped that such an exploration will allow language researchers to gain a more holistic picture of association measure types and their relationships. Although researchers such as Brezina (2018) and Evert (2004, 2007) have provided mathematical comparisons of a few selected association measures, and looked at how high and low-scoring combinations are different between these measures, the added benefit of producing a cluster analysis is that it allows researchers to visualise the mathematical connections between measures and understand empirically how closely their formulas are related, when applied to real-life language data. It should be noted that such an aim is in contrast to much of the association measure literature which chooses a single association measure as the 'best' measure for performing a particular collocation extraction task. In this study, the goal is more focused on retaining distinctly different measures while making the

case for not retaining measures that flag up similar types of collocation and/or collocation properties.

This chapter's aim is achieved by carrying out a cluster analysis on the set of association measures described in Chapter Five (Section 5.5.2). The aim of cluster analysis is to classify objects (in this case, variables), into groups where variables in the same group have similar properties and variables in different groups have dissimilar properties (Pastor, 2010). As outlined by Everitt (1993) and Pastor (2010), cluster analysis can be used as a data reduction technique to reduce large numbers of observations into smaller groups. It is important to point out that group or cluster membership is not fully known beforehand by the researcher. The group or cluster membership emerges from applying the clustering technique (Pastor, 2010).

The chapter proceeds by outlining the collocations that were included in the cluster analysis and how each collocation was allocated a set of association measure scores. The chapter then sets out how the cluster analysis was carried out in terms of the type of cluster analysis and the pre-analysis steps taken with the raw association measure scores. The results of the cluster analysis are then interpreted in light of how the association measures cluster together to show their (dis)similarity. This variability is related back to the theoretical properties of collocation and how these clusters may illuminate different properties of collocation.

6.2 Dependencies in the Cluster Analysis

Since the cluster analysis was based on those dependencies that could be allocated association measure scores, it is important to mention those combinations that were extracted as dependencies but did not feature in the cluster analysis. These combinations included dependencies which were either (a) absent from the MICUSP reference corpus or (b) below threshold in the MICUSP reference corpus. As is standard from past literature (e.g. Paquot, 2018, 2019), below threshold dependencies were those that appeared in the MICUSP corpus less than five times. The below threshold combinations were not included in the analysis as the allocation of an association measure to pairings that appear less than five times in the reference corpus are thought to be unreliable (Evert, 2004; Seretan, 2011). A snapshot of the below threshold and absent combinations are presented in Table 28.

Table 28: Below Threshold and Absent Units

Dependency type	# Dependency types	Below threshold types (%)	Absent types (%)
amod	11,982	23	38
nsubj	3,903	28	46
dobj	9,868	43	33

6.3 Carrying out the Cluster Analysis

After calculating the association measures, the calculation sheet was read into R in .csv file format. In the analysis, each column heading (e.g. t-score) acted as a variable for the cluster analysis and each row with the relevant association measure score acted as a case. By way of illustration for the calculation sheet, Table 29 shows an entry for the known-to-contrast measures of t-score and MI. The calculation sheet included the contingency table components: a, b, c, d and their expected frequencies: a_exp, b_exp, c_exp, d_exp as well as calculations for a+ b, c + d and so forth as dictated by the individual association measure formulas. This allowed each association measure to be calculated in a single calculation sheet.

Table 29: An Entry in the Calculation Sheet

Corpus size	W1	W1 count	W2	W2 count	Collocation count	T score	MI
2,367,652	Ambitious	65	goal	798	5	2.21	6.64

Note: The size of the reference corpus here is the number of words after the text clean up detailed in Chapter Five (Section 5.4.3.3).

The cluster analysis was carried out following advice from multiple sources (e.g. Baayen, 2008; Levshina, 2015 & Gries, 2013b). Hierarchical cluster analysis was carried out because the goal of the analysis was to use the clustering as an initial exploratory method that shows relations between measures as driven by the data rather than any pre-determining theory from the literature or the researcher (Field, Miles & Field, 2012). This differs from non-hierarchical clustering where the researcher may not require such a detailed data analysis and instead pre-determines the number of desired clusters (Manning & Schütze, 1999, cited in Levshina, 2015). Crawley (2013, p.819) sets out the rationale for hierarchical clustering: “The idea behind hierarchical cluster analysis is to show which of a (potentially large) set of samples are

most similar to one another, and to group these similar samples in the same limb of a tree. Groups of samples that are distinctly different are placed in other limbs". This similarity is defined on the basis of the distance between two samples. In this study of clustering variables, this distance will be the distance between pairs of association measures.

The steps involved in carrying out the cluster analysis are summarised in Table 30.

Table 30: Cluster Analysis Steps

Step	Details	Decisions made for association measure data
1. Prepare raw data for analysis	1. Check data for missing values and outliers. 2.Decide on appropriate treatment for these cases. 3.Scale the data to standardize measures across the variables.	1. No missing values found. 2. Outliers allocated a value of the mean for each association measure. 3. Scaling done using 'scale' function in R.
2. Decide on type of clustering	1. Decide on type of clustering method based on desired cluster composition	1. Decided to perform complete or furthest neighbour clustering to create compact clusters.
3. Decide on distance matrix measure and prepare distance matrix	1. Decide on distance measure.	Used Spearman correlations (See Baayen, 2008).
4.Generate dendrogram	1.Generate dendrogram using distance matrix in R	N/A
5.Carry out post hoc validation checks to determine cluster stability	1.Carry out internal validation checks in line with those recommended in the literature	1. Carried out average silhouette width checks.

As the first step in Table 30 shows, a number of pre-analysis steps were carried out on the association measure data. These included exploring each of the association measure variables to identify (i) any missing values and (ii) outliers. Although this step is a standard one for mono and multifactorial analyses, it is particularly important to carry out before a cluster analysis because as Milligan (1980) recognises, the presence of missing data and outliers can significantly change the clustering of the

data. There were no missing values in the data analysis sheet. To determine the extent that the data contained outliers, boxplots for each association measure variable were generated in R and examined. The boxplots helped highlight the number of outliers and their exact location in the data. Outliers were mathematically defined by the boxplots as those values which fell outside the 'whiskers' or ends of the boxplot and are mathematically determined as those values which are outside 1.5 times the interquartile range above the upper quartile and below the lower quartile. In dealing with outliers, there were several possible treatments available (e.g. complete deletion or assign them a particular value). Based on common practice in statistical literature (e.g. See Tabachnick & Fiddell, 2014) and the desire to retain data for the cluster analysis, each outlier was replaced with the mean of that particular association measure variable. This replacement amounted to an average of 17% of the total dependencies.

After this, the numerical data values in each row needed to be scaled. This scaling was necessary to standardise across the different association measures which were based on different scales. If the data were not scaled, potentially some individual values would have contributed an inflated amount to understanding the distance between each pair of variables, when compared in the distance matrix (Baayen, 2008). Following these pre-analysis steps, the cluster analysis was carried out by making a number of decisions in performing Steps 2,3 and 5.

After the scaling, the type of clustering and the type of distance measure used had to be decided. In making these decisions, cluster analysis literature was consulted. The choice of clustering method can result in vastly different clusters forming in the data and as pointed out by Gries (2013b), this decision can be based on what the cluster analysis is being used to show. Gries (2013b) and Tan, Steinbach

and Kumar (2014) specify a number of clustering types. These include single or nearest neighbour, complete linkage or furthest neighbour, and Ward's method of linkage. Each of these are explained and their rationale for selection linked to the type of clusters desired, and the rationale for carrying out a cluster analysis. In single linkage, the similarity of elements X and Y is defined as the minimal distance between any one element of X and any one element of Y so those with the smallest distance would be merged together. Single linkage is deemed to produce long chains of clusters and is therefore not particularly useful for discriminating between clusters. In contrast, in complete linkage or furthest neighbour linkage, the similarity of two objects (e.g. X and Y) is defined as the maximal distance between any one element of X and any one element of Y. This type of linkage tends to form smaller homogenous groups and is a recommended choice if the researcher suspects there are many smaller clusters in the data.

Gries (2013b) points out how Ward's method has a logic similar to ANOVA because it joins those elements whose joining increases the error sum of squares least. Gries (2013b, p.347) states: "For every possible amalgamation, the method computes the sum of squared differences/deviations from the mean of the potential cluster, and then the clustering with the smallest sum of squared deviations is chosen". The ward method is known to generate smaller clusters and has been supported in many applications (Gries, 2013b). Since the primary concern of this analysis is measuring reduction and therefore distinguishing between measures, complete linkage was used to create compact clusters.

After this, the next step was computing the distance matrix. Although, there are a number of options to choose from in choosing a distance measure for example, Izenman (2013) and Tan, Steinbach and Kumar (2014) refer to Euclidean, correlation

and Manhattan measures. Izenman (2013) recommends using a Spearman's correlation matrix because the correlations provide an easily interpretable measure of 'closeness' between pairs of variables and Spearman's correlations are useful for skewed distributions which we often find in linguistic data (see also the approach from Baayen, 2008, p.139). The analysis then used this distance matrix to plot and produce a dendrogram to show the clustering of the whole measure set. The cluster analysis R script is available online[22] with the correlation matrix that forms the distance matrix shown in the form of a heatmap in Figure 6.

Figure 4 shows the resulting dendrogram**.** The dendrogram acts as a tree-like structure that is used to visually represent the (dis)similarity among variables or cases in the dataset. As Field, Miles and Field (2012), Gries (2013b) and Wiechmann (2008) all highlight, the interpretation of such a dendrogram is often not based on stringent objective rules in relation to how many clusters the data form. This is instead more often than not left to the researcher's subjective judgement. Given, this the interpretation of the dendrogram that follows is to an extent based on researcher interpretation. The chapter draws on the definition of a cluster provided by Tan, Steinbach and Kumar (2014,p.493) in that: "A cluster is a set of objects in which each object is closer (or more similar to every other object in the cluster than to any object not in the cluster". The chapter is also influenced by the fact a cluster can be conceptually informed or defined with a cluster simply seen as objects grouped together because they share a particular property or properties. These definitions of a cluster are most pertinent to the study's goal of tapping into aspects of frequency-based collocation properties with the variables in the cluster sharing or drawing on

[22] The R script used to carry out the cluster analysis is available at: https://leemccallum.net/resources/

similar properties. The resulting dendrogram can be interpreted as follows. The dendrogram can be read from the y-axis which represents the linkage distance between the objects. This height symbolises the differences between the variables where the greater the height, the bigger the difference between variables in the cluster analysis. When we look at the structure of the data, we also see several ‘branches’ that connect the measures or objects that are being clustered. These branches vary in length and the longer in length, the greater the difference between the measures/objects (Gries, 2013b). Along the x-axis is simply the variables being clustered together. With this in mind, Wiechmann (2008) explains that objects that are deemed by the clustering to be most similar are allocated to the same cluster and those variables which are not similar are allocated to a different cluster.

In using the cluster analysis to answer the study’s first research question that relates to data reduction and understanding how the association measures may tap into different aspects of collocation properties, Figure 5 helps clarify the relationships between measures. The resulting dendrogram supports a number of relationships between the association measures, with branch height being a determining factor in setting out the degree of (dis)similarity between these measures.

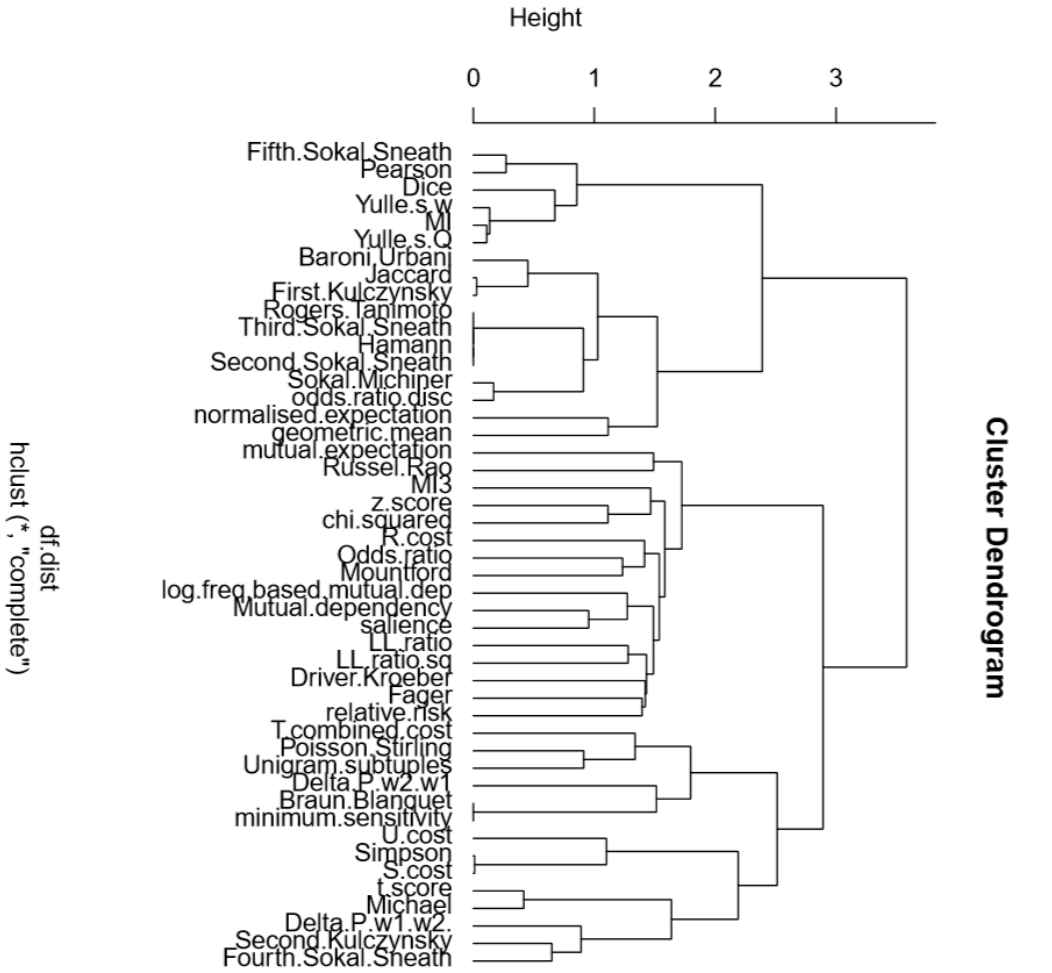


Figure 4: Dendrogram of Association Measures

6.4 Determining the Number of Clusters

The dendrogram visually supports possibly three to six large clusters with a number of sub-clusters apparent within this visual picture. Although a caveat in any cluster analysis work that is appropriate to reinforce here is that like the comments made by Wiechmann (2008), this visual interpretation is laden with subjectivity. Wiechmann (2008, p. 270) notes this difficulty and opts to proceed both pragmatically and cautiously in determining a number of clusters and advises to choose a stable

structure that we can learn from. Before describing the clusters in detail, there are a number of procedures available to check the validation of this visual interpretation. Recently, Brock, Pihur, Datta and Datta (2008) have described a range of possible validation techniques for researchers. These techniques generally fall into two camps: internal and external validation. The former validation takes only the dataset and the clustering solution into consideration in its validation judgement; while the latter makes use of an external dataset or external information about the data into its calculation. In the absence of external information, I opted to carry out internal validation. Brock et al (2008) and Levshina (2015) both recommend using the 'silhouette' function in the cluster package. This function calculates a statistic known as the 'Average Silhouette Width' (ASW) which represents how well-formed the clusters are. Brock et al (2008) explain this as measuring how confident we can be in the clustering solution. Well-formedness means that the members of one cluster are close to one another and far away from the members in the other clusters. The statistic ranges from a value of -1 to +1 with -1 indicating there is no cluster structure present in the data and 1 indicating there is perfect separation of all clusters (Brock et al., 2008). Levshina (2015) presents a rule of thumb to aid interpretation and states that an average silhouette width below 0.2 should be interpreted as a lack of substantial cluster structure in the data. In her own cluster analysis, she achieves an average silhouette width value of around 0.3 and comments that this is above the suggested threshold so we can have some confidence in the clustering.

The average silhouette width statistic was therefore obtained for the three to six clusters that the initial visual interpretation seems to support. These values are presented in Table 31:

Table 31: Average Silhouette Width

Cluster	3	4	5	6
ASW values	0.323	0.328	0.341	0.333

The values suggest that there is greatest support for five clusters with the lower average silhouette width value for six clusters seemingly indicating less stability in forming more clusters. These five clusters are highlighted in Figure 5:

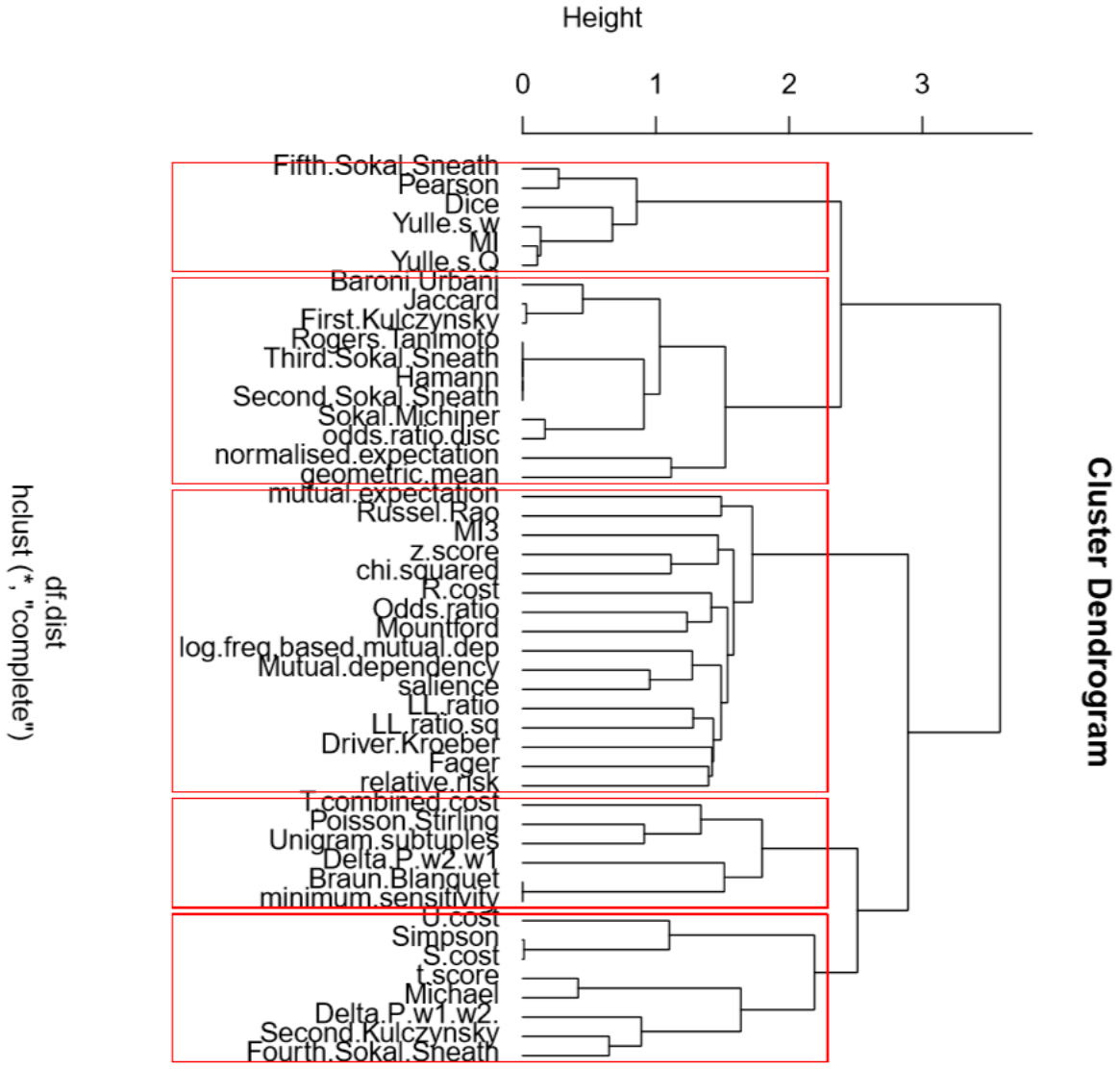


Figure 5: Clusters in the Dendrogram

Before looking at the relationships that exist within clusters, several observations can be made when comparing across the five different clusters. We can see that the low branch height appearing in clusters one, two, four and five (heights range from below 1.0 to below 2.0) indicates that the measures are most similar i.e. have the least degree of difference whereas the measures in cluster three have higher branch height indicating there is greater difference between the measures in this cluster. The implications of these observations for the goal of measure reduction will now be discussed for each of the five broad clusters.

6.4.1 Cluster One

Reading Figure 5 from the left-hand side, cluster one contains the following association measures: Fifth Sokal Sneath, Pearson, Dice, Yulle's W, MI and Yulle's Q. In cluster one, there appears to be some difference between the Fifth Sokal Sneath and Pearson compared with the Yulle's measures and the MI. For example, the lack of height difference between MI and Yulle's ω and Yulle's Q seems to suggest these are more similar than any of the other measures. With regards to branch height, the height appearing close to 0 indicates that the difference between the Yulle's ω, Yulle's Q and MI is negligible. There appears to be a degree of difference between the Dice measure and the other five measures because its branch height is higher signalling more of a difference between this measure and the more tightly bound other five.

6.4.2 Cluster Two

Cluster Two contains the following association measures: Baroni Urbani, Jaccard, First Kulczynsky, Rogers Tanimoto, Third Sokal Sneath, Hamann, Second Sokal Sneath, Sokal Michiner, Odds ratio disc, Normalised expectation and the Geometric mean. When comparing cluster one with cluster two, in Figure 5, we see greater branch height to suggest measures in this group share greater difference between pairs than those in cluster 1. However, like cluster one, there is still evidence of collinearity between pairs of measures. The geometric mean appears to be the most distinct from the other measures given its branch height is highest. When we look at the pairing up of ‘Sokal Michiner’ and ‘Odds ratio’, the lower branch height compared to the ‘geometric mean’ suggests evidence of collinearity between the measures. Like the geometric mean there appears to be a degree of uniqueness to the measure of normalised expectation given the higher branch height compared to others in the cluster. In contrast, what stands out in cluster two is the low height connecting the other measures. For example, the low height between Sokal-Michiner and Odds ratio disc and the visibly flat connections between Hamann, the Second Sokal Sneath, Rogers Tanimoto, and the Third Sokal Sneath; and later the Jaccard and ‘First Kulcznsky. It appears that like the geometric mean, normalised expectation and Baroni-Urbani maintain some degree of difference when compared with these other measures. Like the measures in cluster one, these measures pairing up at low height suggest high degrees of collinearity between measures.

6.4.3 Cluster Three

Cluster Three contains the following association measures: Mutual expectation, Russel Rao, MI^3 , zscore, chi-sqaured, R cost, Odds ratio, Mountford, Log frequency based mutual dependency, Mutual dependency, Salience, Log-likelihood ratio, Log-likelihood ratio squared, Driver Kroeber, Fager, and Relative Risk. In contrast to the first two clusters, cluster three's starting height and branch length appear to indicate that the paired-up measures share the greatest variation. This seems to indicate that these measures share the least collinearity and are in fact the most likely cluster to contain measures tapping into different collocation information than those measures in the other clusters. This variation is fairly consistent with only the links between mutual dependency and salience having notably shorter branch height than the other pairings.

6.4.4 Cluster Four

Cluster Four contains the following association measures: T combined cost, Poisson Stirling, Unigram subtuples, Delta P w1 I w2, Braun Blanquet, and Minimum Sensitivity. When we compare the measure sets of clusters one-three with cluster four, we see greater variation in the similarity properties of the measures because the starting branch height is higher meaning the measures are notably different in cluster four from those in the first three clusters. However, when we examine the cluster membership, the measures appear to have varying degrees of dissimilarity. There appears to be little difference in 'Braun Blanquet' and 'Minimum Sensitivity' while the 'T Combined Cost' measure appears distinct from others in the cluster and the 'Poisson Stirling' measure and the 'Unigram Subtuples' measure appear to have some degree of difference as again indicated by branch height.

6.4.5 Cluster Five

Cluster Five contains the following association measures: U cost, Simpson, S cost, T score, Michael, Delta P w1 l w2, Second Kulczynsky, and the Fourth Sokal Sneath. Cluster five contains similar patterns to those we see in cluster four with greater branch height differences. The flat height between Simpson and S cost indicate high degrees of collinearity. Overall, the greater height between the pairs of measures here seems to indicate that although the measures cluster together, there is noted variation in what their values appear to be highlighting/focused on. This seems to support the literature on how association measures have been grouped together so clusters four and five contain measures that are heuristic, hybrid or mathematical transformations of each other. It is also worth pointing out that clusters four and five again help highlight how there are differences between the theoretical groupings that Evert (2004) and Pecina (2010) have used.

In this respect, we have identified five large clusters of measures. Given that the goal of this study was to reduce the number of association measures into a grouping or set that could capture as much information about the notion of collocation as possible, the following seems to hold true:

- Measures in clusters one, two, four and five contain measures with varying degrees of collinearity (as indicated by their lack of branch height).
- Measures in cluster three seem to share the greatest variation and therefore least collinearity.

However, outside these visual observations, there is still a need to use this information to help inform measure use/retention decisions. The next section of the chapter looks

at the distance between members of each cluster and uses this information to make judgements on the retention of specific measures within and across clusters so as to tap into different collocation properties.

6.5 Justifying Measure Retention

Although the dendrogram is a useful visualisation of the relationships between association measures, it is also helpful to examine this visualisation with established association measure guidance in mind. Therefore, to help determine the numerical properties that guide the clustering, a heatmap was generated to show the relationships that exist between and within the clusters estimated in the dendrogram. The heatmap is shown in Figure 6[23]. The heatmap is a visual representation of the distance between the association measures. It can be interpreted as showing the strongest correlations (closest distances) in red, with purple representing weak to approaching moderate correlations and the blank white spaces showing little correlation (greatest distance apart) between measure pairs.

With language learning in mind, Gablasova et al (2017) and Schneider (2018) are two primary studies that develop a rationale for measure selection. Gablasova et al (2017) advise that measure selection should be led or informed by three criteria: (1) the mathematical computation of the measure, (2) the scale it operates on and (3) its practical effect (i.e. what combinations get flagged up and what combinations get downgraded). While Schneider (2018) encourages users of association measures to choose measures for language learning that are most able to tap into human judgement or the psycholinguistic 'priming' or probability we subconsciously use as

[23] A clearer visualisation of this map can be found at: https://leemccallum.net/resources/

we mentally “pair up” words for use. The particularly comprehensive advice from Gablasova et al (2017) can be applied to the clustering here.

The first point worth being reminded of is that the clusters represent similarity and therefore in each cluster the aim is to decide to retain a single measure or a small sub-set of cluster members that appear representative enough of the cluster and its underlying collocation properties. The sections that follow draw on both the available collocation literature, the empirical dendrogram and heatmap and this advice from Gablasova et al (2017) to help determine particularly useful and distinct measure patterns that arise from the clusters.

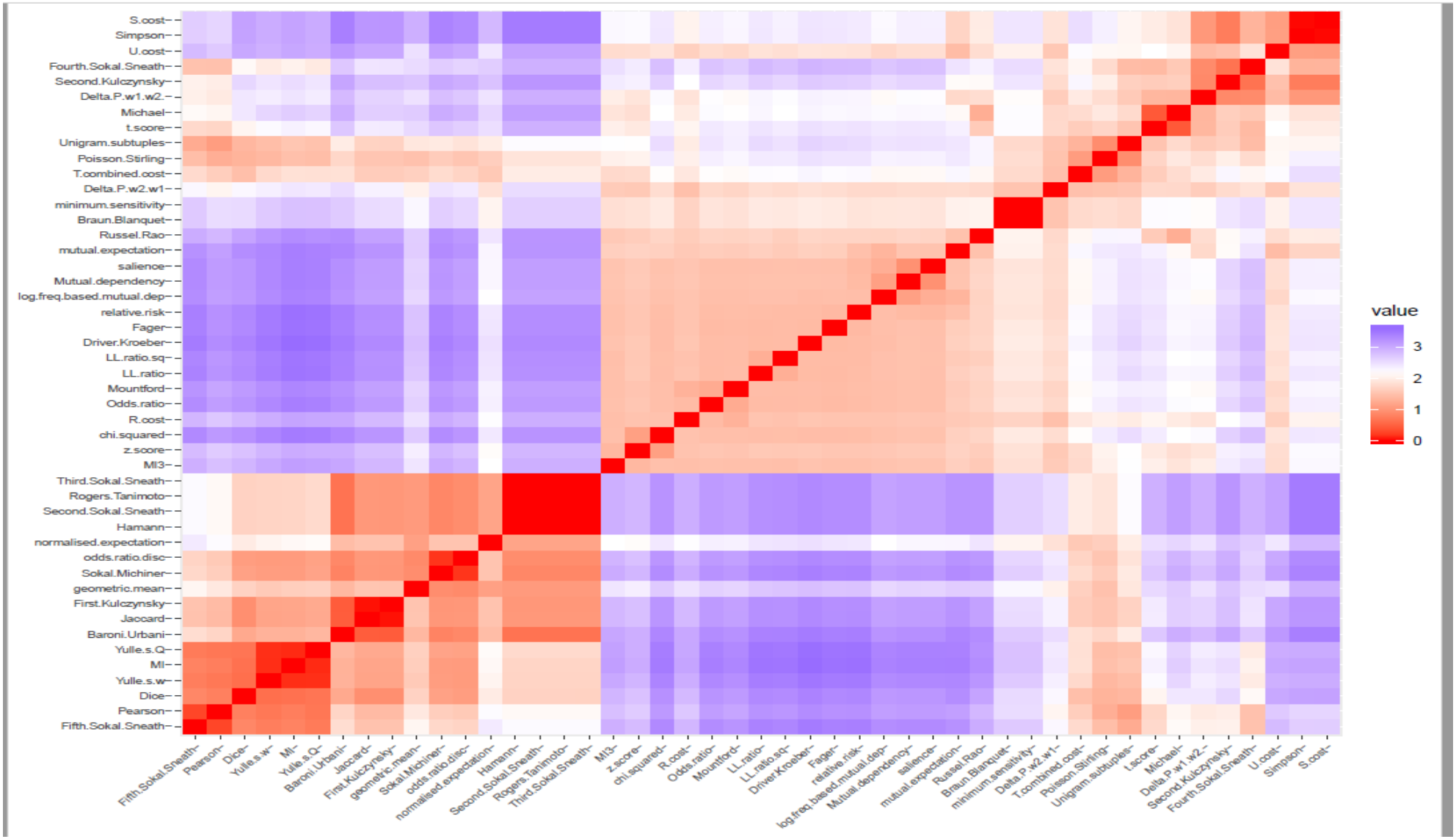


Figure 6: Heatmap Showing Relationships across Association Measures

6.5.1 Cluster One

With respect to Gablasova et al's (2017) first criteria of mathematical computation, we can appreciate that when we examine the heatmap of correlations between the individual association measures, it shows that there is clear evidence of strong collinearity between MI and the Yulle's measures. Equally, Dice seems to indicate in the opposite respect a greater degree of difference although it must be highlighted that the correlations between measures in this cluster are very high overall (r values range from .802 to .998). Given this high degree of collinearity and that we know the MI has been shown to tap into exclusivity this leads to a conclusion that the measures in this cluster may flag up similar highly exclusive pairings. In making a decision to retain measures in this first cluster, it is sensible to consider scale as Gablasova et al (2017) recommend but it is also important to consider the ease with which we can interpret this scale and make use of it in the focus on collocation properties. Therefore, it seems logical to consider scale and practical effect together when deciding between cluster members.

Although Gablasova et al (2017) appear to strongly offer support for the scaling of the Log Dice, there is an arguably greater rationale for supporting the retention of the MI. Despite the fact the MI does not operate on an easily theoretically interpretable minimum or maximum scale, extensive exploration with learner language data has shown that an MI ≥ 3 holds promise for robust collocation use whereby these combinations are more likely to also be grammatically sound pairings that appear in native texts and in the case of combinations with a score greater than 3, there is ample and emerging evidence that these combinations are in fact examples of specialised, discipline, or genre specific language (Durrant & Schmitt, 2009; Granger & Bestgen,

2014; Bestgen & Granger, 2014; Paquot, 2018, 2019). Therefore, the measure most appropriate to retain from cluster one is the MI.

6.5.2 Cluster Two

What is perhaps most noticeable in cluster two is the lack of association measures that have a concrete background in language learning, teaching or assessment research. As the literature review (Chapter Four, Section 4.4) highlighted, many coefficient association measures, like the ones here, remain underexplored in learner language. Therefore, in making decisions about measure retention in this cluster, the criteria of mathematical computation and practical effects are guiding principles. Starting with measures that appear highly correlated, Figure 6's heatmap and referring back to the dendrogram in Figure 5, we can see high correlation (or close distance if we look back at the dendrogram) between Jaccard and First Kulczynsky; and between Rogers-Tanimoto, Third Sokal Sneath, Hamann and Second Sokal Sneath. The correlations between these measures exceed r=.80 and at the same time, these measures actually show evidence of high collinearity with other measures in the other clusters (in some cases this reaches r=≥.73). For example, there are strong correlations (r = ≥.74) between the Third Sokal Sneath and Dice and similarly strong correlations between Baroni-Urbani and the Jaccard and the First Kulczynsky. Following other feature-grade literature (e.g., Kyle & Crossley, 2016, p.17) and general statistics literature (e.g., Tabachnick & Fidell, 2014, pp. 122-123) on what constitutes collinearity (r=≥.70 -.90), we can see that the retention of all of these measures would more than likely mean retaining measures that capture very similar properties of collocation. The value of the cluster analysis appear clear here. Without such an analysis, the relationships between the many coefficients mentioned in the literature

would remain unclear. In relation to this point, this cluster, more than the others, seems to offer counter evidence to the claim from Schneider (2018). As evidenced in the literature review, Schneider (2018) stated in his discussion of Delta P and transitional probability that they shared high collinearity (r= >.90), and it was rare to encounter highly correlating association measures. In the case of this cluster, there seems to be evidence that this is in fact a possibility when we examine their respective linking in the cluster analysis.

For cluster two, the measure we can be most confident about is the geometric mean. It is a good representative of the cluster because of its strong correlations with other cluster two members (r values peak at 0.687), and its correlation to other association measures in different clusters is lower when compared to other cluster two members. For example, other cluster two members have higher correlations to cluster one members, which approach multicollinearity, for example the correlation between the odds ratio disc and the MI reaches 0.80); while the correlations between cluster one measures and the geometric mean and cluster one measures only peaks at 0.64 and goes as low as 0.40 with the LogDice, below the threshold for multicollinearity. This means the geometric mean can be safely retained as different *enough* from cluster one members and other clusters three to five.

6.5.3 Cluster Three

In looking at cluster three's relationships and applying Gablasova et al's (2017) criteria, there is a need to draw more on comments made in the hypothesis testing literature since many of the measures in this cluster are measures of significance that Evert (2004) and Pecina (2005, 2010) as well as language learning literature (e.g. Durrant & Schmitt, 2009; Bestgen & Granger, 2014; Granger & Bestgen, 2014; Paquot, 2018, 2019) have referred to.

As pointed out in Section 6.4.3, the branch heights indicate that measures in this cluster are not tightly bound together. The original correlations shown in the heatmap in Figure 6 also support the largely independent nature of the measures in this particular cluster. Many of the measures in this cluster originate within the hypothesis testing category hence the degree of some correlation between them. When the degrees of collinearity are explored from the heatmap, the strongest correlation exists between the measures of mutual expectation and Russel Rao (r=.18). All other measures have very low to no correlation. This provides us with some important understandings of association measures as they act independently and together. Gries and Durrant (2020) have previously noted that association measures are manipulations or regressions of each other, and while this is indeed true for much of the observations in other clusters, cluster 3 seems to indicate that within these pools of measures, there are groups which share weak correlation.

Referring to the criteria from Gablasova et al (2017), very few of the measures in this cluster have concrete use in language learning studies. There is no concrete evidence of use of Mutual expectation, Russel Rao, R cost, Mountford, Log frequency based mutual dependency, Mutual Dependency, Salience, Driver Kroeber, Fager or Relative Risk in language learning studies. In large-scale computational studies that have focused on identifying a 'best' measure for identifying collocations of a particular type (e.g. Pecina, 2005, Pecina, 2010), these measures have not emerged as best performing candidates either.

The most prominent and perhaps 'enduring' measures in language learning have been framed around those which are significance test measures. For example, the z-score and the odds ratio have featured in language studies (Gries & Durrant, 2020). However, Evert (2007) highlights the scale difficulty with the odds ratio in that

the measure's scale is difficult to interpret intuitively. The most pragmatic alternative 'sound' measures that appear to be appropriately grounded in language learning studies is the Log-likelihood ratio and its seemingly more mathematically robust squared alternative: Loglikelihood ratio squared. On inspection in the heatmap and the dendrogram, there appears to be support for these two measures.

With regards to Gablasova et al's (2017) criteria, there is a pragmatic argument to be made that either of the loglikelihood ratios (loglikelihood ratio and loglikelihood ratio squared) are pragmatic robust choices. Both are robust measures because they operate on an interpretable scale, have some history in language learning and have received support from large-scale computational studies as meaningful, robust and sound association measures (Evert, 2004, p.21; Pecina, 2010). Given that there has been a tendency to favour mathematically robust and less-bias measures that do not overinflate language data, the squared loglikelihood ratio seems to be the most pragmatic choice and is therefore the measure retained in cluster 3. However, it should be highlighted here that, as a whole, cluster 3 presents us with a number of unexplored association measures that may, with further exploration tap into potentially different properties of collocation that other measures in the cluster do not capture in a similar way. It is also worth considering the implications of the lack of relationships in this cluster with those found in other clusters and how this influences the picture of association measures as a whole.

6.5.4 Cluster Four

Like in cluster three, none of the measures have particularly strong connections to language learning literature. We see a weak clustering in cluster four with sub-groups appearing between the Braun Blanquet and the Minimum Sensitivity (r=1.00) and these two measures share high correlations with the measures Poisson Stirling and Unigram Subtuples. However, we see only weak correlations between the Delta P w2|w1 and the T-combined cost in relation to these measures. This suggests that the differing principles of calculation that underpin these measures mean they do not cluster together as well as the more stable clustering in clusters one and two. For example, their highest correlations with other measures reach r=.354 between the T combined cost and Poisson Stirling; and Delta P w2 l w1 has a maximum correlation of r=.212 with Unigram subtuples.

When looking at cluster four in relation to the other three previous clusters, there is evidence that the measures of the T-combined cost and the Delta P w2|w1 are able to better highlight different properties of collocation than the measures of Braun Blanquet, Minimum Sensitivity, Unigram subtuples and Poisson Stirling because these measures have higher correlations to measures in clusters one to three.

At the same time, there is little supporting information to rationalise the selection of the remaining measures of Unigram subtuples and Poisson Stirling because these have not been examined in language learning or indeed gained any support from extensive experimental trials on collocation data. Therefore, the most logical conclusion to be drawn is that the measure Delta P w2 l w1 should be retained for Cluster four.

6.5.5 Cluster Five

Taken as a whole, cluster five contains a similar sub-group structure to cluster four in that some measure pairs are highly correlated, while others have loose clustering. Like the highly correlated measures in clusters one, two and four, the measures of Simpson and S cost are also highly correlated. These two measures are also highly correlated (r=. 899, r=.895) with the Second Kulczynsky and the Fourth Sokal Sneath (r=.725, r=.728). The same high correlation can be found between the t-score and Michael (r=.952). Although, these two measures correlate highly with each other, they do not share exceedingly high correlations with their other respective cluster members (r=.002, r=.655).

Since the t-score has the more prominent background in language teaching and we know it has the potential to highlight highly frequent word combinations that are used generically across multiple genres (Durrant & Schmitt, 2009; Granger & Bestgen, 2014), this measure along with the seemingly distinct Delta P word 1 and word 2 should be retained in this cluster.

The decisions made here on retention have been made by looking at the mathematical relations between the measures, their scales and also their previous history in language learning. It is also important to look holistically at the cluster analysis and reflect back on the methodological clustering choices made at the start of Chapter 6. The choice of furthest neighbour clustering was made so as to produce tight clusters. We can see then that we have a spectrum of clusters, with clusters two, three, four and five further away from the measures and underlying principle of 'exclusivity' that cluster one is clearly based upon. With reference to clusters four and five we can see a sparse picture of clustering with these measures furthest from those in cluster one, but not necessarily always strongly clustered together as pairs in their

respective cluster. There is evidence, then, that this clustering corroborates the theoretical work of Brezina (2018) and Evert (2004, 2007) who have made reference to the properties that are flagged up by the association measures, operating along a cline or continuum. This continuum goes from exclusivity at one end to more of a focus on significance testing/raw frequency measures at the other end. The visual cluster supports this work and also highlights how the measures of the Delta P continue to be distinct from this potentially dichotomous picture, given their low correlation/cluster ties with other non-directional association measures.

Overall, the literature review helped highlight the broad nature and groupings of association measures with Evert (2004) and Pecina (2005, 2010) grouping the bulk of these measures together under similar families. However, the cluster analysis here helps shed light on how family members are related to each other, and how they are distinct from others. This visual picture is not obtained to dispute these previous groupings; instead it sought to act as a complement that highlights how some of the measure pairings in the established literature are related to or distinct from others.

6.6 Summary

This chapter looked at the relationships between different association measures spread out across the broad groupings that emanate from past literature. The chapter used the information garnered from the cluster analysis to build a rationale for explaining which measures would be particularly relevant in capturing different aspects of collocation properties. A final decision was reached to retain the following measures: MI, Geometric mean, Log-likelihood Ratio^2, Delta P word 2 word 1, T-score, and Delta P word 1 word 2.

The penultimate chapter of this thesis takes forward these measures into the analysis of student writing and how these measures and their properties have a relationship with students' writing score grades.

Chapter Seven: Mixed-effects Modelling

7.1 Introduction

This chapter brings together many of the theoretical and empirical claims made throughout this study in relation to the relationship between measures of collocation and writing quality scores. As outlined in the literature review and methodology chapters, there is a pressing need to measure such a relationship via the use of mixed effects modelling. This modelling has an advantage over monofactorial methods because it can more robustly measure this relationship via the introduction of random effects that account for experimental variation in the corpus. With this measurement claim in mind, this penultimate chapter has the central goal of empirically showing how the relationship between collocations and writing quality can be measured more reliably by accounting for such variation.

The chapter proceeds as follows. First, the inclusion of collocation dependencies is discussed in Section 7.2. Then, the variables of interest for the modelling are described in Section 7.3 and following this the modelling process is undertaken and interpreted in the remainder of the chapter for the 'Final_Grade' outcome variable. A standard ordinal regression model is used in Section 7.4 to answer research question two (**RQ 2:** To what extent do measures of collocation have a relationship with writing quality?) by looking at both the effects of the association measures (answering research question **RQ 2.1:** When a set of these association measures are selected, to what extent do they have a relationship with writing quality?) and the fixed effects of language status and task (answering research question **RQ 2.2:** To what extent do the fixed effects of task and language status also have a

relationship with writing quality?). The findings of this modelling process are discussed and an explanation as to why this model is “anti-conservative” or presents a simplistic view of relationships is then put forward. The sections that follow this anti-conservative model, detail how a more robust understanding can be gained via the introduction of random effects which account for the complex nature of the sampling corpus. This mixed effects modelling therefore answers research question 2.3 (**RQ 2.3:** To what extent do these relationships vary when the modelling process considers the random effects of individual rater and/or individual student?). The overall findings from the mixed effects modelling are then related back to pertinent collocation and feature-writing grade literature and the FYC context itself.

7.2 The Corpus Make Up and Collocations in the Modelling Process

For calculating diversity measures, all dependency types were counted irrespective of their status in the reference corpus. This meant that frequency counts for amod, nsubj and dobj types were based on the raw counts of frequency types generated from an R script and therefore counted the variation in grammatical dependencies since the measure includes both collocations and non-collocations.

For calculating association measures that acted as a proxy for understanding the sophistication of collocations, this was done only for units that appeared five or more times in the MICUSP corpus. This meant that the allocation of an association measure score was only carried out after excluding below threshold types (those types appearing less than 5 times in MICUSP) and those types which were completely absent. It seems pertinent here to point out that while L2 studies have included below threshold and absent types as legitimate predictors or correlates of writing quality grades (e.g., Granger & Bestgen, 2014, Bestgen & Granger, 2014; Paquot, 2019), I chose not to include these measures. This decision was made after inspecting the

types and also with a consideration of the reference corpus chosen. Many of the below threshold and absent types were simply found to fall into this category because of task differences between the FYC corpus and the MICUSP corpus, that is to say, they were not below threshold or absent because they were grammatically or collocationally incorrect (e.g., in the case of Granger & Bestgen, 2014) and so it seemed unlikely that the status of these units would be related back to telling us something about their relationship to writing quality grades. Therefore, the decision helped highlight that the status of such below threshold and absent units will be dependent on the reference corpus chosen. In the case of Granger and Bestgen's (2014) study, they employ COCA and their rationale that below threshold or absent types will have a relevance to writing quality is warranted since the scope for language use is much broader than in the case of the study here, which uses another specialised corpus as its reference corpus. This meant that it was taken as natural to encounter perfectly grammatical types which are simply below threshold or absent in MICUSP because of the nature of the corpus, the writing it contains and ultimately its smaller size when compared to general corpora such as COCA or the BNC.

7.3 Variables in the Traditional Regression Model

In the modelling process, the holistic grades act as the dependent variable. As a dependent variable, the holistic grades were treated as ordinal in nature. This was because the grades could be ordered from lowest to highest (7 – 15 points). This decision had an important influence on the type of regression model chosen. Since, the dependent variable was ordinal with multiple categories (to represent the different grade bands), the model type chosen was ordinal logistic regression (see Christensen, 2018; Finch, Bolin & Kelley, 2014; Winter, 2020 for an overview on ordinal data). Christensen and Brockhoff (2013) highlight how ordinal data are commonplace across many disciplines where humans are used as measurement instruments. He includes the examples of school gradings and consumer ratings of preference as examples of ordinal data. Liu (2016) is another who highlights the benefits of such a model type and regards them as suitable for ordinal outcome variables which are categorical in nature with ranks or orders. This includes examples such as students' socioeconomic status ordered low to high; children's proficiency in early reading scored from level 0 to 5; or survey data with responses from strongly disagree to strongly agree. The same logic can be applied to that of essay grades where there is some kind of evaluation taking place where raters order essays with some rated better than others. Therefore, this kind of modelling approach seems to be the best logical option for the FYC dataset where final grades are awarded on a scale of whether or not competencies are 'low emerging', 'emerging, 'high emerging', 'low developing', 'developing', 'high developing', 'low mastering', 'mastering' and 'high mastering' (with a range of 7 – 15 points awarded for these grade levels). Christensen and Brockhoff (2013) advocates

the use of *cumulative link models*[24] for their ability to preserve the ordering of categorical variables. This is also a rationale for not following binomial regression since the variable 'Final_Grade' actually has 9 levels or categories, or multinomial logistic regression which does not retain the ordering information for the variables (Christensen, 2018).

The cumulative links model represents the effect of the independent variable on the dependent variable as logit odds which are the mathematically calculated odds of the dependent variable having an effect on the independent variable (O'Connell, 2006). These odds are often transformed into an easier to interpret odds ratio which Levshina (2015) notes is also an effect size or in simple terms an estimate of the effect of this regression relationship.

To illustrate the measurement effects for research question two, a fixed effects cumulative links model was created with the dependency measures acting as fixed effects. Winter (2020, p. 236) summarises the status of a fixed effects variable as a variable that is repeatable and constant across experiments. In this sense, then, a fixed effects variable is one where we could repeat a study on differences in a variable (e.g., gender) by collecting data with new females and males: while the individual participants vary (and their individual differences are a source of 'random' variation on the data), the effect of gender can be tested again and again with new samples. Much like this rationale, the variables of 'task' and 'language status' can also be repeatedly tested and are assumed to have a predictable, non-idiosyncratic influence on the response that could be tested with new writers. The same logic can also be applied to the measures of collocation in that they could also be tested with new writers.

[24] These models are also known as proportional odds models (Liu, 2016).

For each learner text, a mean association score was calculated for each dependency type so as to give a mean score for each association measure for each individual text. For the diversity measures, the Root TTR for each dependency type (amod, nsubj and dobj) was also calculated for each individual text. For the fixed effects of 'Task' and 'Language_status', two levels of coding were used to distinguish the tasks from each other, and the language status of the writers from each other. A more detailed explanation of this coding and how it is handled in the modelling process is set out in Section 7.4.1.3.

7.4 Collocations and Writing Quality: Fixed Effects

The first step in this modelling process was to centre and standardise each of the linguistic variables. Centring involves shifting a variable's mean to 0 by subtracting the mean from the variable. Standardisation is a transformation that involves converting variables into a standard scale. This initial step is recommended in modelling literature for both theoretical and practical interpretation reasons. First, from a theoretical perspective, Winter (2020) and Finch et al (2014) support the use of centring and standardisation because they convert the independent variables into a scale of standard units and this helps in making variables comparable, when, for example, assessing the impact of multiple predictors. Second from a practical perspective, such procedures are also thought to later avoid convergence issues (Winter, 2020, p. 266). The data were therefore centred and transformed using the centring procedures and Box Cox transformations in Gries (2013b).

After this, the next step in the modelling process was to generate a fixed-effects model that included all possible predictors. The model was generated with the package 'ordinal' in R and used the function 'clm' which generated a cumulative links model. The 'ordinal' package is able to handle data which is ordered i.e. in the case of grade levels, these can be ordered highest to lowest or lowest to highest and more importantly, the package and its functions are able to deal with multiple levels of order or different categories of the ordinal variable (Christensen, 2018). This differs from other types of R packages and embedded types of regression which manage binomial regression (e.g., when there are two categories for the outcome variable) and multinomial regression (which can handle more than two categories but does not preserve the order of the categories) (e.g., See Finch et al., 2014). Therefore, the ordinal package was chosen for its flexibility. The package also allows users to include a variety of random effects including that of nested or crossed variables (Christensen, 2018).

The full model was then examined and trimmed down following standard modelling procedure from Zuur et al (2009) and Gries (2015) so that non-significant predictors were removed and only retained if their removal influenced the significance of other predictors (Gries, 2015). At each stage of this trimming, I removed one variable at a time (in order of highest p-value since this indicated variables that were not significant) and compared the generated model to the previous one. I used the 'anova' function to compare models and I inspected the AIC (Akaike Information Criteria) value as a check of goodness of fit for the models. (Gries, 2015[25]). The AIC

[25] Gries (2015) uses this method of model evaluation and trimming while pointing out that no measure of evaluation is without criticism in statistics however his approach does follow other core modelling literature (e.g., Zuur et al., 2009; Brezina, 2018).

value works on the understanding that the most parsimonious model is the one that explains as much variation in the dependent variable as possible with as few predictors as possible (Brezina, 2018). The lowest AIC value indicates the most parsimonious model. A close inspection of the R script[26] can be used to explain what takes place during such modelling. The R script shown in Figure 7 can be explained as fitting a model for fixed effects where the first line dictates that everything on the right-hand side of the tilde ('~') is an independent variable (e.g., an association measure) and the left-hand side of the '~' contains the dependent variable (i.e. the grade score).

```
model.fixed<-clm(Final_Grade ~ 1 + meanMIamod + meangmeanamod +
meanLLR2amod + meandeltapw2w1amod + meandeltapw1w2amod +
meantscoreamod + meanMInsubj + meangmeannsubj + meanLLR2nsubj +
meandeltapw2w1nsubj + meandeltapw1w2nsubj + meantscorensubj + meanMIdobj +
meangmeandobj +  meanLLR2dobj + meandeltapw2w1dobj + meandeltapw1w2dobj +
meantscoredobj + amodRTTR + nsubjRTTR + dobjRTTR + Task + Language_status,
data=data)

summary(model.fixed)
```

Figure 7: Code for Full Model

A snapshot of this trimming procedure is shown in Figure 8. Figure 8 shows the first trimming between the full model and the next model with the first predictor removed. The lower AIC value shows that the 'model.fixed.optimal' is a better fit and the p-value comparison (p=0.9592) indicates that the full model 'model.fixed' is not significantly better than the reduced 'model.fixed.optimal' model meaning the new

[26] The full R script is available at: https://leemccallum.net/resources/

model to take forward and trim further is the ‘model.fixed.optimal’ one (following procedures in Gries, 2015).

```
link: threshold:
model.fixed.optimal logit flexible
model.fixed         logit flexible

                    no.par    AIC  logLik LR.stat df
model.fixed.optimal     30 3613.9 -1776.9
model.fixed             31 3615.9 -1776.9  0.0026  1
                    Pr(>Chisq)
model.fixed.optimal
model.fixed             0.9592
```

Figure 8: Snapshot of Model Comparison

7.4.1 Interpreting the 'Final_Grade' Fixed Effects Model

The trimmed model output is shown in Figure 9. The analysis that follows is divided into providing an explanation of how the output should be interpreted (Section 7.4.1.1) and then this interpretation is divided into a discussion of the collocation measures (Section 7.4.1.2) and then that of the role of task and language status (Section 7.4.1.3).

```
link  threshold   nobs logLik   AIC
 logit flexible   879  -1782.79 3599.58


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
meanLLR2amod        -0.18697    0.06119  -3.056  0.00224**
meanMInsubj          0.30811    0.11683   2.637  0.00836**
meandeltapw2w1nsubj -0.19198    0.08377  -2.292  0.02193*
meantscorensubj     -0.22794    0.09622  -2.369  0.01784*
meanMIdobj           0.18298    0.06866   2.665  0.00770**
meandeltapw1w2dobj  -0.14217    0.06525  -2.179  0.02934*
amodRTTR             0.17683    0.07097   2.491  0.01272*
nsubjRTTR            0.13775    0.06988   1.971  0.04868*
Language_status2     0.34860    0.14288   2.440  0.01470*

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
Estimate Std. Error z value
7|8   -3.21486    0.17838 -18.023
8|9   -2.53887    0.13444 -18.884
9|10  -1.89751    0.10665 -17.791
10|11 -1.21965    0.08820 -13.828
11|12 -0.51674    0.07816  -6.611
12|13  0.12376    0.07598   1.629
13|14  1.07193    0.08428  12.719
14|15  2.02393    0.10766  18.799
```

Figure 9: The Fixed Effects Model

7.4.1.1 Interpreting the log odds and odds ratios

Winter (2020) notes that operating solely on the log odds scale can perplex readers and so transforms the results into odds ratios like Liu (2016). This transformation is done by exponentiating the coefficient to give a more intuitive easier to interpret scale. The logit odds to odds ratio conversions are shown in Table 32 and the threshold coefficients are shown in Table 33. These odds were obtained via the code 'round(exp(coef(model.fixed.optimal13)),3)' in R where the original logit odds were exponentiated to transform them into odds ratios and then the code 'round(exp(confint(model.fixed.optimal13,type = "Wald")),3)' obtained the 95% confidence intervals with both the exponentiated and confidence intervals rounded to three decimal places. The log odds or odds ratio expresses the odds of an increase or decrease in the dependent variable (e.g., grade) with a one unit increase of the independent variable (e.g., measure of collocation), when all other variables are held constant. The threshold coefficients represent the cumulative nature of the model. They highlight that the model essentially performs a series of level comparisons and the cumulative model is the cumulation of those series of comparisons i.e. the overall model (Liu, 2016).

Liu (2016) advises to interpret the logit odds as follows:

- When the logit odds sign is positive, the converted odds ratio is > 1. This means that the odds of being beyond a particular category increases for a one-unit increase in the continuous predictor variable.
- When the logit odds sign is negative, the converted odds ratio is < 1. This means that the odds of being beyond a particular category decrease for a one unit increase in the continuous predictor variable.
- When the logit coefficient equals 0, the odds ratio equals 1. This means that there is no relationship between the predictor and the odds, so there is no change in the odds when values of the continuous predictors change.

At the simplest level of understanding, such a model is testing the odds of grades moving up or down the grading scale relative to either an increase in units for continuous predictors (e.g., the collocation measures), or when either level of a categorical predictor (e.g., language status). Section 7.4.1.2 and 7.4.1.3 that follow are therefore based on the interpretation of the coefficients and confidence intervals in Tables 32 and 33.

Table 32: Converted Odds Ratio Coefficients for Final Grade Model

Predictor variable	Logit coefficients	Std. Error	z value	Pr(>|z|)	Odds Ratio	95% Confidence Intervals	
						2.5%	97.5%
meanLLR2amod	-0.187	0.061	-3.056	0.002	0.829	0.736	0.935
meanMInsubj	0.308	0.117	2.637	0.008	1.361	1.082	1.711
meandeltapw2w1nsubj	-0.192	0.084	-2.292	0.022	0.825	0.700	0.973
meantscorensubj	-0.228	0.096	-2.369	0.018	0.796	0.659	0.961
meanMIdobj	0.183	0.069	2.665	0.008	1.201	1.050	1.374
meandeltapw1w2dobj	-0.142	0.065	-2.179	0.029	0.867	0.763	0.986
amodRTTR	0.177	0.071	2.491	0.013	1.193	1.038	1.372
nsubjRTTR	0.138	0.070	1.971	0.049	1.148	1.001	1.316
Language_status2	0.349	0.143	2.44	0.015	1.417	1.071	1.875

Table 33: Converted Threshold Cut-Off Points for Final Grade Model

Threshold coefficient	Logit coefficients	Odds Ratio	95% CIs	
			2.5%	97.5%
7|8	-3.215	0.040	0.028	0.057
8|9	-2.539	0.079	0.061	0.103
9|10	-1.898	0.150	0.122	0.185
10|11	-1.220	0.295	0.248	0.351
11|12	-0.517	0.596	0.512	0.695
12|13	0.124	1.132	0.975	1.313
13|14	1.072	2.921	2.476	3.446
14|15	2.024	7.568	6.128	9.346

7.4.1.2 The collocation measures

The fixed effects model in Figure 9 and its summarised information in Tables 32 and 33 indicates that nine predictors in total were significant predictors of the outcome variable of 'Final_Grade' : eight of these predictors were collocation measures and the other was the non-linguistic predictor of 'Language_status2'. The eight significant collocation measures included four positive coefficient collocation predictors: the logit coefficient for MeanMI nsubj ('MeanMInsubj') (β = .308, SE = .117, z =2.637, p=.008); the logit coefficient for meanMI dobj ('MeanMIdobj') (β = .183, SE = .069, z =2.665, p=.008); the logit coefficient for amod diversity ('amod RTTR') (β = .177, SE = .071, z =2.491, p=.013) and the logit coefficient for nsubj diversity ('nsubj RTTR') (β = .138, SE = .070, z =1.971, p=.049). In addition to four negative coefficient predictors: the logit coefficient for Mean MI LLR2amod ('MeanLLR2amod') (β = -.187, SE = .061, z = - 3.056, p=.002); the logit coefficient for Mean Delta P w1w2 nsubj ('Meandeltapw2w1nsubj') (β = -.192, SE = .084, z = - 2.292, p=.022); the logit coefficient for Mean tscore nsubj ('Meantscorensubj') (β = -.228, SE = .096, z = - 2.369, p=.018); and the logit coefficient for Mean Delta P w1w2 dobj (Meandeltapw1w2dobj) (β = -.142, SE = .065, z = - 2.179, p=.029).

In terms of the odds ratio (OR), the odds of being beyond a particular 'Final_Grade' score, were 1.361 times greater with a one unit increase in MeanMInsubj. The odds of being beyond a particular 'Final_Grade' score, were 1.201 times greater with a one unit increase in MeanMIdobj. The odds of being beyond a particular 'Final_Grade' score, were 1.193 times greater with a one unit increase in amod RTTR. The odds of being beyond a particular 'Final_Grade' score, were 1.148 times greater with a one unit increase in nsubj RTTR. In other words, texts containing

higher ‘MeanMInsubj’, ‘MeanMIdobj’ ‘amod RTTR’ and ‘nsubj RTTR’ profiles have higher odds of being awarded a higher final grade.

Those collocation predictors with negative logit coefficients can be interpreted as follows. In terms of the odds ratio (OR), the odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.829 for a one unit increase in ‘MeanLLR2amod’. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.825 for a one unit increase in ‘meandeltapw2w1nsubj’. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.796 for a one unit increase in ‘meantscorensubj’. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.867 for a one unit increase in ‘meandeltapw1w2dobj’. In other words, texts containing higher ‘MeanLLR2amod’, ‘meandeltapw2w1nsubj’, ‘meantscorensubj’ and ‘meandeltapw1w2nsubj’ profiles have lower odds of being awarded a higher final grade.

From the original full model, the non-significant measures that were removed from the model included a number of measures that have previously been found to have some relationship with grade score. For example, while Paquot (2019) found that the mean MI amod was a significant predictor, in the present study this was not the case: instead the Mean MI nsubj and Mean MI dobj were among the significant predictors. None of the Geometric mean (gmean) measures performed well in the modelling process and neither did many of the Log-Likelihood Ratio squared measures (‘LLR2’) as evidenced by their absence from the final fixed effects model. This indicates that these measures do not have a statistically significant relationship with grade and may suggest that these particular measures do not flag up word combinations that raters are paying attention to.

7.4.1.3 Task and language status

For the interpretation of the variables 'Task' and 'Language_status' in the fixed effects model, it is important to understand their coding procedure in R. Since there were two levels for task, I coded this as 1 = Task 1, and 2 = Task 2. Similarly, since there were 2 levels for 'Language_status', I coded these as 1 = native speaker and 2 = non-native speaker.

In the model, the algorithm in R chooses one of these levels to act as a reference level and uses this reference level as a comparison with the other levels[27]. In the model output, the reference level will not be shown, but the level compared to it will be (e.g., see Winter, 2020, p. 184; & Liu, 2016). This means that for the model output in Figure 9, 'Language_status1' (native speaker) is the reference level.

In the modelling process the predictor of 'Task' was not a significant predictor of final grade. This means that there was no significant evidence that the odds of a higher grade being awarded for either task 1 or task 2. This finding seems to counter some of the literature which has also considered this source of grade variation. For example, Quellmalz et al (1982), Ruth and Murphy (1988) and more recently Barkaoui (2008) have all pointed out this source of variation in grade allocation.

When the two speaker types are compared, the original logit coefficient is positive meaning that 'Language_status2' or non-native speaker status increases the odds of being beyond a particular grade level. In terms of the odds ratio (OR), the odds of being beyond a particular grade level are 1.417 times greater when the speaker is a non-native speaker of English. This result indicates that non-native writers have higher odds of receiving a higher grade when all other predictors in the model are held constant. This result may help shed light on sources of rater bias. However, it is also

[27] This is normally the first numerical level or alphabetical (Levshina, 2015).

possible that non-native writers pay more attention to their writing because they are aware of their supposed lack of writing ability and therefore work harder to improve. The statistical result obtained here is interesting in light of the FYC rater make-up (see discussion of this make up in Chapter Two (Section 2.5, p.32), and later in this chapter (p. 244) and in concluding remarks in Chapter Eight (Section N, p.255). Future FYC research could unpack this finding further by interviewing FYC raters about their particular perceptions of L1 and L2 writers' texts and how their perceptions translate into making judgements using the FYC task rubrics. This would help understand the statistical results further.

7.4.1.4 Confidence in the Fixed Model

Confidence in the model obtained can be obtained by inspecting the 95% confidence intervals (CIs). Such an inspection of Table 33 shows that:

- For the 'MeanMInsubj' measure, the odds ratio is estimated to be 1.361 and the confidence interval range [1.082 – 1.711] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the 'MeanMIdobj' measure, the odds ratio is estimated to be 1.201 and the confidence interval range [1.050 – 1.374] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the 'amod RTTR' measure, the odds ratio is estimated to be 1.193 and the confidence interval range [1.038 – 1.372] indicates that we can be 95%

confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.

- For the ‘nsubj RTTR’ measure, the odds ratio is estimated to be 1.148 and the confidence interval range [1.001 – 1.316] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the ‘Language_status’ measure, the odds ratio is estimated to be 1.417 and the confidence interval range [1.071 – 1.875] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.

For the negative coefficients, the confidence intervals, can be interpreted as:

- For the ‘LLR2amod measure’, the odds ratio is estimated to be 0.829 and the confidence interval range [0.736 – 0.935] indicates that we can be 95% confident that the true odds ratio covers in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.
- For the ‘meandeltapw2w1nsubj’, the odds ratio is estimated to be 0.825 and the confidence interval range [0.700 – 0.973] indicates that we can be 95% confident that the true odds ratio covers in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.
- For the ‘meantscorensubj’, the odds ratio is estimated to be 0.796 and the confidence interval range [0.659 – 0.961] indicates that we can be 95% confident that the true odds ratio lies in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.

- For the ‘meandeltapw1w2dobj’ measure, the odds ratio is estimated to be 0.867 and the confidence interval range [0.763 – 0.986] indicates that we can be 95% confident that the true odds ratio covers this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.

It is also worth commenting on the threshold coefficients. Liu (2016) highlights that the threshold coefficients are essentially a series of binary regressions across two levels of the ordinal outcome variable. This can be seen in Table 33, where a series of comparisons are made between grade levels 7|8, 8|9, 9|10 until the highest grades of 14|15. Essentially these threshold pair comparisons are a reflection of the cumulative model. In Table 33, it is important to illuminate the different logit and odds ratios that appear across these comparisons. In Table 33, we see that the logit odds are negative with small odds for those grades from 7-12. However, there is marked shift in direction at comparisons between 12|13, 13|14 and 14|15 whereby the logit odds and the odds ratio become positive with largest odd increases especially between grades 13|14 and 14|15. This could indicate that these predictors have more chance of increasing grade only at these higher levels, and that there is less chance of the predictors increasing low grade levels.

7.5 Corpus Structure: Theoretical and Mathematical Dependency

The next step in the modelling process was to establish if it was necessary to include consideration of the hierarchical structure of the corpus in the modelling process. The fixed-effects model in Figure 9 assumes that one student wrote each essay and that an independent rater rated one essay only. Therefore, there is an assumption that the data points are independent. However, institutional practice on the FYC programme

means that a single rater grades essays from multiple classes and because of data collection the same student may contribute to both tasks meaning there is more than one text per student. Therefore, the random effects modelling looks at how much these sampling issues are able to explain grade variation.

A close examination of the structure of the FYC dataset indicates that raters cross-over between classes and therefore rater and class are crossed variables in the dataset (see Baayen, Davidson & Bates, 2008; Bates, 2010 for an overview of crossed variables). In R, this can be seen by cross-tabulation between the variables of 'Rater_ID' and 'class_id' whereby raters appear in more than one class. A subset of this structure is shown in Table 34. For example, focusing on the rater with Rater_ID '4649', Table 34 shows that this rater grades essays across multiple classes as opposed to only marking essays in one class. This data structure can be thought of as each rater having multiple membership in the variable of 'class_id'. This data structure therefore introduces dependency into the dataset because grades awarded by the same rater are likely to share a degree of similar variance over a situation where each rater is only solely responsible for a single essay.

Table 34: Cross Tabulation for Rater ID and Class ID

	Class_id											
Rater_ID	10254	10258	10260	10273	10277	10279	10960	11465	11466	11470	11508	12098
2780	0	0	0	0	0	0	0	0	0	0	0	0
2921	0	0	0	0	0	0	0	0	0	0	0	0
3435	0	0	0	0	0	0	0	0	0	0	0	0
3516	0	0	0	0	0	0	0	0	0	0	0	0
4649	0	0	0	0	0	0	10	0	9	6	0	0
4988	12	0	0	0	0	12	0	0	0	0	9	0
5315	0	0	0	0	0	0	0	0	0	0	0	0
5570	0	0	0	0	0	0	0	0	0	0	0	0
5634	0	0	0	0	0	0	0	0	0	0	0	0
7254	0	0	0	0	0	0	0	0	0	0	0	10

It is important here to point out that this crossed structure differs from a nested structure whereby if class was nested within rater, only one rater would appear in a single class (Baayen et al., 2008). To take account of the rating situation at FYC a 'full random' model was generated, consisting of student + class + rater to reflect the cross-over between rater and class and the fact that students may submit more than one text across classes. Although Luke (2011) states that the primary justification for mixed-effects modelling arises from theoretically knowing there is dependency in the data, it is possible to mathematically measure this in the dataset. The Intraclass Correlation (ICC) can be calculated to mathematically represent this dependency or clustering. The ICC is defined as "the proportion of the variance explained by the grouping structure in the population" (Hox 2002, p. 15). This value ranges from 0 to 1 with 0 indicating that there is no dependency in the structure that could explain any of the variance; and 1 indicating that there is substantial dependency in the dataset and that this grouping structure is responsible for the variance.

To determine if the clustering within rater, class and student was something that needed to be accounted for the ICC was calculated using the formula:

ICC = $\frac{\sigma_u^2}{\sigma_{u\ +\sigma^2}^2}$ where σ_u^2 is the variance of the random effects and σ^2 is the residual variance, which, in cumulative models is assumed to equal π^2 / 3.29 (O'Connell, 2010; Liu, 2016). The resulting ICC value from this model equalled 0.409 (rounded to three decimal places) indicated the effect of rater should be included as a variable in the modelling process (see guidance from Liu, 2016). However, to determine the optimal random effects structure I also generated a series of simpler models consisting of simply rater, rater plus student, rater plus class, student, and student plus class to

determine which structure would account for the most variance but have the lowest AIC value and therefore produce the more parsimonious model. From this modelling process, I compared the random models with each other, again using the 'anova' function. The comparisons indicated that the rater only model was a more robust model, as evidenced by its lower AIC value (O'Connell, 2006). The comparisons between these models is shown in Table 35. For these reasons, the optimal random structure was judged to be simply 'Rater_ID' shown fully in Figure 10.

Table 35: Random Model Comparisons

Random model comparison	Model components	AIC value	Loglikelihood values	Total variance
1	rater, class and student	3490.11	-1734.06	1.348
2	rater and class	3489.72	-1734.86	0.979
3	rater and student	3488.11	-1734.06	1.348
4	rater	3487.72	-1734.86	0.978
5	student and class	3543.89	-1761.94	1.309
6	student	3633.88	-1807.94	0.379

```
Cumulative Link Mixed Model
fitted with the Laplace approximation

formula: Final_Grade ~ 1 + (1 | Rater_ID)
data:    data

 link  threshold nobs logLik   AIC
 logit flexible  879  -1734.86 3487.72

Random effects:
 Groups   Name        Variance Std.Dev.
 Rater_ID (Intercept) 0.9784   0.9891
Number of groups:  Rater_ID 45

No Coefficients

Threshold coefficients:
      Estimate Std. Error z value
7|8   -3.63478    0.23749 -15.305
8|9   -2.94576    0.20603 -14.298
9|10  -2.28103    0.18826 -12.116
10|11 -1.55720    0.17721  -8.787
11|12 -0.77637    0.17098  -4.541
12|13 -0.03683    0.16885  -0.218
13|14  1.07170    0.17272   6.205
14|15  2.16388    0.18753  11.539
```

Figure 10: Random Effects Model with Rater

7.6 Collocations and Writing Quality: Mixed-Effects

Based on the preliminary exploration of the corpus structure, a mixed-effects cumulative links model was generated. In this manner, it is perhaps helpful to think of these two optimal random and fixed models as two halves joining together in the mixed model to account for as much variance as possible. The final mixed effects model is shown in Figure 11.

```
Cumulative Link Mixed Model fitted with the Laplace approximation

formula:
Final_Grade ~ 1 + meanLLR2amod + meanMInsubj + meandeltapw2w1nsubj +
    meantscorensubj + meanMIdobj + meandeltapw1w2dobj + amodRTTR +
    nsubjRTTR + Language_status + (1 | Rater_ID)
data:    data

 link  threshold nobs logLik   AIC
 logit flexible  879  -1714.52 3465.05

Random effects:
 Groups   Name        Variance Std.Dev.
 Rater_ID (Intercept) 0.9349   0.9669
Number of groups:  Rater_ID 45

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
meanLLR2amod        -0.13858    0.06323  -2.192  0.02841*
meanMInsubj          0.34035    0.12173   2.796  0.00518**
meandeltapw2w1nsubj -0.15766    0.08590  -1.835  0.06647.
meantscorensubj     -0.25293    0.09998  -2.530  0.01141*
meanMIdobj           0.12881    0.07168   1.797  0.07236.
meandeltapw1w2dobj  -0.09537    0.06853  -1.392  0.06402.
amodRTTR             0.16184    0.07336   2.206  0.02738*
nsubjRTTR            0.17869    0.07503   2.381  0.01725*
Language_status2     0.16253    0.14811   1.097  0.07248.

---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
      Estimate Std. Error z value
7|8   -3.69543    0.24017 -15.387
8|9   -3.00267    0.20920 -14.353
9|10  -2.33204    0.19169 -12.166
10|11 -1.59395    0.18041  -8.835
11|12 -0.79064    0.17354  -4.556
12|13 -0.02986    0.17096  -0.175
13|14  1.10864    0.17464   6.348
14|15  2.22638    0.18972  11.735
```

Figure 11: The Final Mixed Model

Figure 11 highlights that those predictors from the fixed effects model that reached significance continue to be significant predictors although the variable of ‘language _status2’ is only marginally significant in the mixed-effects model and the variables’, logit odds and odds ratio (OR) vary. The sections that follow interpret the mixed model independently and then compare it to the fixed effects model to appreciate their respective differences.

7.6.1 Interpreting the Random Effects

The random effect of rater indicates that there is considerable variation between raters with the variation between each individual rater just under one grade point (variance =0.9349, SD = 0.9669).

7.6.2 Interpreting the Collocation Measures

The mixed effects model in Figure 11 and its summarised information in Tables 36 and 37 indicates that nine predictors in total were significant predictors on the outcome variable of 'Final_Grade' : eight of these predictors were collocation measures and the other predictor was 'Language_status2'. The eight significant collocation measures included four positive coefficient collocation predictors: the logit coefficient for MeanMI nsubj ('MeanMInsubj') (β = .340, SE = .122, z =2.796, p=.005); the logit coefficient for the mean MI dobj ('MeanMIdobj') (β = .129, SE = .072, z =1.797, p=.072); the logit coefficient for amod diversity ('amod RTTR') (β = .162, SE = .073, z =2.206, p=.027); the logit coefficient for nsubj diversity ('nsubj RTTR') (β = .179, SE = .075, z =2.381, p=.017). As the only non-linguistic predictor, language status ('Language_status2') was also significant (β = .163, SE = .148, z = 1.097, p=.072).

The remaining four negative coefficient predictors were: the logit coefficient for Mean MI LLR2amod ('MeanLLR2amod') (β = -.139, SE = .063, z = - 2.192, p=.028); the logit coefficient for Mean Delta P w2 w1 nsubj ('Meandeltapw2w1nsubj') (β = -.158, SE = .086, z = - -1.835, p=.066); the logit coefficient for Mean tscore nsubj ('Meantscorensubj') (β = -.253, SE = .100, z = - 2.530, p=.011); and the logit coefficient

for Mean Delta P w1w2 dobj (‘Meandeltapw1w2dobj’) (β = -.095, SE = .069, z = -1.392, p=.064).

In terms of the odds ratio (OR), for the positive coefficients, the odds of being beyond a particular ‘Final_Grade’ score, were 1.405 times greater with a one unit increase in MeanMInsubj. The odds of being beyond a particular ‘Final_Grade’ score, were 1.137 times greater with a one unit increase in the MeanMI dobj. The odds of being beyond a particular ‘Final_Grade’ score, were 1.176 times greater with a one unit increase in amod RTTR. The odds of being beyond a particular ‘Final_Grade’ score, were 1.196 times greater with a one unit increase in nsubj RTTR. In other words, the use of higher ‘MeanMInsubj’, ‘MeanMIdobj’, ‘amod RTTR’ and ‘nsubj RTTR’ increase the odds of texts being awarded a higher final grade score.

Those collocation predictors with negative logit coefficients can be interpreted as follows. In terms of the odds ratio (OR), the odds of being at or beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.871 for a one unit increase in MeanLLR2amod. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.854 for a one unit increase in ‘meandeltapw2w1nsubj’. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.777 for a one unit increase in ‘meantscorensubj’. The odds of being beyond a particular ‘Final_Grade’ score, decrease by a factor of 0.909 for a one unit increase in ‘meandeltapw1w2dobj’.

These results indicate that the odds of being beyond a particular grade level decrease with a one unit increase in the predictors of ‘MeanLLR2amod’, ‘meandeltapw2w1nsubj’, ‘meantscorensubj’ and ‘meandeltapw1w2dobj’.

7.6.3 Interpreting Language Status

In the mixed effects model, ‘language_status2’ indicates that the odds of non-native writers being beyond a particular grade level were 1.176 times greater than native writers. It is important to recognise the importance of this contextual finding and place it side-by-side with the findings of those that were positive linguistic coefficients. These results seem to indicate that in the case of the positive coefficients, there is a similar size of odds for grade increase with language status and both amod and nsubj diversity with the odds higher for language status, helping put into perspective two things. First, that there is a tendency for grade increases to be shaped by multiple linguistic and non-linguistic variables, and in this case the non-linguistic predictor has slightly higher odds of grade increases than the linguistic predictors.

Table 36: Converted Odds Ratios for Mixed-Effects Model with Final Grade

Predictor variable	Logit coefficients	Std. Error	z value	Pr(>|z|)	Odds Ratio	95% Confidence Intervals	
						2.5%	97.5%
meanLLR2amod	-0.139	0.063	-2.192	0.028	0.871	0.769	0.985
meanMInsubj	0.340	0.122	2.796	0.005	1.405	1.107	1.784
meandeltapw2w1nsubj	-0.158	0.086	-1.835	0.066	0.854	0.722	1.011
meantscorensubj	-0.253	0.100	-2.530	0.011	0.777	0.638	0.945
meanMIdobj	0.129	0.072	1.797	0.072	1.137	0.988	1.309
meandeltapw1w2dobj	-0.095	0.069	-1.392	0.064	0.909	0.795	1.040
amodRTTR	0.162	0.073	2.206	0.027	1.176	1.018	1.357
nsubjRTTR	0.179	0.075	2.381	0.017	1.196	1.032	1.385
Language_status2	0.163	0.148	1.097	0.072	1.176	0.880	1.573

Table 37: Converted Thresholds for Mixed-Effects Model with Final Grade

Threshold coefficient	Logit coefficients	Odds Ratio	95% CIs	
			2.5%	97.5%
7|8	-3.695	0.025	0.016	0.04
8|9	-3.003	0.050	0.033	0.075
9|10	-2.332	0.097	0.067	0.141
10|11	-1.594	0.203	0.143	0.289
11|12	-0.791	0.454	0.323	0.637
12|13	-0.030	0.971	0.694	1.357
13|14	1.109	3.030	2.152	4.267
14|15	2.226	9.266	6.389	13.44

7.6.3.1. Confidence in the Mixed Model

In terms of how confident we can be that these results reflect the 'true' odds of 'Final_Grade' increasing or decreasing, an examination of the 95% confidence intervals in Table 37 indicates that:

- For the 'MeanMInsubj' measure, the odds ratio is estimated to be 1.405 and the confidence interval range [1.107 – 1.784] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the 'MeanMIdobj' measure, the odds ratio is estimated to be 1.137 and the confidence interval range [0.988 – 1.309] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the amod RTTR measure, the odds ratio is estimated to be 1.176 and the confidence interval range [1.018 – 1.357] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.
- For the nsubj RTTR measure, the odds ratio is estimated to be 1.196 and the confidence interval range [1.032 – 1.385] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.

- For the ‘Language_status’ measure, the odds ratio is estimated to be 1.176 and the confidence interval range [0.880 – 1.573] indicates that we can be 95% confident that the true odds ratio covers this range. In other words, if the model was repeatedly estimated, the true odds would lie in this range, 95% of the time.

For the negative coefficients, the confidence intervals, can be interpreted as:

- For the ‘LLR2amod measure’, the odds ratio is estimated to be 0.871 and the confidence interval range [0.769 – 0.985] indicates that we can be 95% confident that the true odds ratio covers in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.
- For the ‘meandeltapw2w1nsubj’, the odds ratio is estimated to be 0.854 and the confidence interval range [0.722 – 1.011] indicates that we can be 95% confident that the true odds ratio covers in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.
- For the ‘meantscorensubj’, the odds ratio is estimated to be 0.777 and the confidence interval range [0.638 – 0.945] indicates that we can be 95% confident that the true odds ratio lies in this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.
- For the ‘meandeltapw1w2dobj’ measure, the odds ratio is estimated to be 0.909 and the confidence interval range [0.795 – 1.040] indicates that we can be 95% confident that the true odds ratio covers this range . In other words, if the model was estimated again, the true odds would lie in this range, 95% of the time.

Like in the fixed effects model, we can also examine the threshold coefficients that make up the cumulative main model. In Table 36, it is important to illuminate the different logit and odds ratios that appear across these comparisons. In Table 37, we see the same patterns from the fixed model in that the logit odds are negative for grade comparisons 7-13, the odds of the predictors, increasing grades decrease. There is a noticeable shift in direction at the two highest grade levels, that of comparisons between 13 and 14, and 14 and 15 whereby the logit odds and the odds ratio become positive. This appears to reiterate that, as found in the fixed model, there predictors have an increased chance of increasing grade only at these higher levels.

7.6.4 Overall Interpretation of the Mixed Model

On the whole, the mixed-effects model shows that the predictors of 'MeanMInsubj', 'MeanMIdobj', amod and nsubj diversity dependencies have greater odds of increasing final grade scores. In contrast, the predictors of meanLLR2amod','meandeltapw2w1nsubj','meantscorensubj',and 'meandeltapw1w2dobj' have greater odds of decreasing final grade scores. To some extent this mirrors the findings from Paquot (2019) who also found support for the MI as did other studies (e.g., Durrant & Schmitt, 2009; Bestgen & Granger, 2014; Granger & Bestgen, 2014). However, in the case of this FYC context, support for the MI is for nsubj and dobj dependencies rather than amod dependencies. It appears that in the FYC context, rather than being more likely to award higher grades for more sophisticated amod dependencies like Paquot (2019), the odds of FYC raters awarding higher grades to more diverse amod

and nsubj use is greater. It is also worth bearing in mind that since these diversity measures contain both dependencies that would reach or not reach collocation status according to past MI thresholds, then it seems it is more likely that raters tend to be overall influenced by the number of dependency types and to a lesser extent their average sophistication.

7.7 Model Evaluations

To evaluate the goodness-of-fit for the mixed effects models, there are a number of statistics available for evaluation (Field, Miles and Field, 2012; Gries, 2015; Levshina, 2015; Liu, 2016). Although traditional linear modelling literature has often cited the R-squared value to give an estimate of how much variance in grade is explained by a particular model, such an estimate is not recommended for this kind of logistic regression. It can be noted that several studies have used a 'Pseudo-R squared' value to replicate the linear modelling R-squared (e.g., Liu, 2016), however, this is not widely applied by those who have used it without noted caution.

Levshina (2015) comments on the use of Pseudo R^2 as a measure of goodness-of-fit. Like many others (e.g., Field, Miles and Field 2012; Liu, 2016), she notes that in logistic regression models, the value of the R^2 is often lower than the R^2 in linear regression, even if the quality of models is comparable (e.g., see comments from Hosmer & Lemeshow, 2000, p. 167). Equally, Levshina (2015) notes that the Pseudo R^2 is less conceptually clear than the R^2 from linear regression[28]. In the ordinal package used in this study, the computation of a Pseudo R^2 value is not supported for the reason that the literature has not borne

[28] It is often misinterpreted as being exactly the same concept as the linear R^2 however this is disputed by scholars who have made direct criticisms of the statistic (e.g, Hosmer & Lemeshow, 2000).

out a suitable or reliable equivalent for the R^2 supported in linear regression (Christensen, personal communication, 2020).

Instead of this measure, the models created in this chapter can simply be evaluated by looking at the AIC values.

Table 38: Model Statistics for Final_Grade Models

Dependent Variable	Model	AIC values	Loglikelihood values
Final_Grade	Fixed	3599.58	1782.79
	Random	3487.72	1734.86
	Mixed	3465.05	1714.52

Table 38 reinforces the idea that modelling work needs to take account of both fixed and random variables to produce a more robust account of the relationships and predictors at play in the sampling corpus. Table 38 also highlights that although the fixed model has the highest Loglikelihood value, it is at the expense of being a less parsimonious model.

7.8 Discussion

The predictors that significantly inform understandings of collocation-writing quality relationships indicate that there is a higher chance that raters assign higher grade scores to texts which contain higher MI scoring nsubj and dobj dependencies. The final mixed effects model also appears to indicate that texts with more amod and nsubj diversity also appear to be more likely to receive higher scores from raters, and that L2 writers are also marginally more likely to receive a higher grade from raters. In contrast, the variables relating to the log-

likelihood squared measure for amod dependencies and the mean Delta P w2w1 nsubj, and mean Delta P w1w2 dobj and the mean tscore for nsubj dependencies appear less likely to have a chance of texts receiving a higher grade.

As part of the discussion around explaining these results further qualitative analysis was carried out. Since the mean of each association score was used in the modelling, I looked at the use of high and low scoring dependency combinations so as to connect and understand the use of high and low scoring dependencies which lie behind high or low mean association measure scores per text. These examples are shown in Table 39, while low scoring MI nsubj dependencies are shown in Table 40.

Taken together, Tables 39 and 40 offer an important understanding of high and low scoring MI units. Table 39 highlights how common high scoring combinations comprise of combinations which have clearly identifiable textual functions. Combinations such as 'paper__nn_:_focus__vb' and other 'paper + verb' combinations are used to set out what the essay intends to do. These functions are shown in examples [15] – [18], taken from different FYC essays.

[15] This **paper** will **focus** on the two most popular opinions, for and against players linked to steroids being treated the dame for consideration for the Hall of Fame; then find common ground between the two sides so that there may be a compromise.

[16] This **paper** will **focus** on different views on air pollution between China National Chemical Corporation in Beijing and local citizens.

[17] This **paper** will specifically **focus** on efforts for animal rights in the entertainment industry, that being animals in zoos.

[18] To sum up, this **paper** would **discuss** how social media affect higher education.

A close examination of low scoring MI combinations in Table 40 highlights how these combinations comprise of pairings of common lexical verbs such as ‘be’ and ‘have’, which can have many other word partners and in the FYC corpus, they do not appear to perform clear textual/rhetorical functions, give an indication of genre or discipline, as examples [19] – [21] indicate:

[19] The parties should not call for complete abolishment of standardized tests in the college application system: the main **goal** of the reform should **be** to shift the focus of college applications from standardized tests to a more well-rounded form of student evaluation.

[20] Uber’s main **goal** would **be** to achieve better public relations due to the political firestorm caused by taxi companies suing Uber for the variety of reasons above.

[21] The **possibility** in order to help treat the veterans for post-traumatic stress disorder would **be** to have them go through a mental evaluation.

The examples in Table 39 and 40 also highlight an important observation relating to the use of high and low scoring combinations to fulfil rhetorical functions. In texts with higher average MI scores, the rhetorical functions are performed through the use of higher scoring MI dependencies while in lower mean scoring texts, the rhetorical functions are still performed, but they are performed using combinations with lower MI scores, and are thus less exclusive pairings. For example, the use of combinations such as those including the verb ‘be’ have low MI scores because they are not exclusive pairings and are thus taken to be less specialised in their use. These findings partly align with Paquot (2018,2019) who also commented that combinations which receive higher MI scores tend to have

greater identifiable disciplinary/genre functions while those with lower scores tend to comprise of word pairs which are less exclusive and can therefore have multiple possible partners.

Table 39:High Scoring MI Nsubj Dependencies

Nsubj dependency	MI score
paper__nn_:_focus__vb	4.41
paper__nn_:_discuss__vb	4.30
evidence__nn_:_support__vb	4.28
paper__nn_:_address__vb	4.25
lesson__nn_:_provide__vb	3.49

Table 40: Low Scoring MI Nsubj Dependencies

Nsubj dependency	MI score
goal__nn_:_be__vb	-1.89
possibility__nn_:_be__vb	-1.75
explanation__nn_:_be__vb	-1.34
study__nn_:_provide__vb	0.66
use__nn_:_make__vb	1.02

Tables 41 and 42 help build a picture of high and low-scoring dobj dependencies. Table 41 shows a range of high-scoring MI dobj dependencies; of which many of these also perform textual functions (e.g., bridge__vb_:_gap__nn) or are specific to the topics being discussed, as highlighted in examples [22] – [26]. These examples indicate how high scoring MI dobj dependencies are used to put forward how solutions may aid a compromise between stakeholders (i.e. completing the task in module ENC 1102, e.g., example [22]); or are used to engage with the perceived advantages of a particular initiative (e.g. example [23]), or highlight what a particular essay will

achieve (e.g., example [24) or present different stakeholder initiatives (e.g., examples [25] and [26]):

[22] If not through social change then through legal changes, such as the bill in Maryland, can **bridge** the **gap** between the two sides on this issue.

[23] The National Education Association withholds an ongoing commitment to **bridge** the **gap** between implementation and standards, through curriculum, training and support.

[24] Not only this, it will also **shed light** on work-life balance, lack of funds and survival in recessions and how it is confronted by corporate owners to be successful in the competitive market.

[25] The main goal of this mission will be to **shed light** on the icy potential habitability, but it could also search for signs of alien life.

[26] They also work to improve the legal system to prevent these false convictions and to **shed light** on the inhumanity of the death penalty.

Table 41: High Scoring MI Dobj Dependencies

Dobj dependencies	MI score
bridge__vb_:_gap__nn	11.23
shed__vb_:_light__nn	11.05
take__vb_:_course__nn	10.75
steal__vb_:_money__nn	9.79
prevent__vb_:_fraud__nn	9.40

Table 42: Low Scoring MI Dobj Dependencies

Dobj dependencies	MI score
have__vb_:_value__nn	-1.19
have__vb_:_role__nn	-1.00
have__vb_:_experience__nn	-0.78
have__vb_:_ability__nn	-0.74
have__vb_:_time__nn	-0.60

In contrast, for texts with lower average MI scores overall, many of the low scoring MI dobj dependencies in Table 42 are found to be being used to perform a variety of rhetorical functions, as the examples [27] – [31] show, including discussing a compromise between stakeholders (example [31]), setting out the implications of a compromise (e.g., example [29]), or bringing the reader into the text by posing a question (e.g,. example [27]), however, the dependencies comprise of highly frequent lexical verbs (e.g., have) which can have many other alternative partnering nouns meaning the dependencies here are less exclusive:

[27] Do all people have rights to **have** this **value**?

[28] This allows the reader to feel a sense of community regarding individuals across the globe, along with a sense of community between people and nature, all of which Monsanto is claiming to **have** a **role** in making happen.

[29] If they both choose to work together in the near future, the life of student athletes would be the best it has ever been and athletes would **have** a more meaningful college **experience**.

[30] Most see this as very ethical as the animals do not **have** an **ability** to help themselves get better.

[31] This will allow them to **have** more **time** to focus on the classes they are taking to ensure they get a good grade, more free tim, less stress, and it will

allow them to be involved in more organizations and possibly take leadership roles or internships to prepare for their future.

This contrast again aligns with Paquot (2018, 2019) who comments that many low-scoring combinations comprise of ‘nuclear’ or basic units of vocabulary which are neither specific to academic or disciplinary writing.

Looking at dependencies which had less chance of increasing grade, the list of LLR2amod combinations in Table 43 shows that high scoring combinations and Table 44 shows the low-scoring combinations. Table 43’s combinations have a clear connection to the topics of the essays; while Table 44’s combinations are less specific and comprise of more general or common words such as: ‘many’, ‘good’, ‘other’ and ‘different’.

Table 43: High Scoring LLR2 Amod Dependencies

Amod dependency	**LLR^2 score**
High__jj_:_school__nn	1930.80
Domestic__jj_:_violence__nn	1331.51
Renewable__jj_:_energy__nn	1325.69
Social__jj_:_movement__nn	1155.75
Great__jj_:_deal__nn	870.40

Table 44: Low Scoring LLR2 Amod Dependencies

Amod dependency	**LLR^2 score**
Other__jj_:_health__nn	7.87
Good__jj_:_student__nn	7.58
Many__jj_:_group__nn	7.56
Different__jj_:_system__nn	7.33
Other__jj_:_process__nn	7.03

Although, interpretation and explanation of why these dependencies have a lesser chance of yielding grade increases is only be limited to an inspection of such ranked dependencies, it is worth considering that these dependencies may yield less chance because they are used frequently and that the repeated use may be perceived less favourably and, in some way, viewed as the 'taken for granted' way to discuss these respective topics. However, such findings are in great need of further qualitative analyses in order to unpack this relationship further.

Similarly, the same can be said for Delta P measures. Little previous research helps guide an evaluation of these combinations since in the first and second language literature, and the wider computational linguistics literature, use of this association measure is only just starting to emerge. However, when ranked from highest to lowest, we can observe some interesting contrasts. High scoring Delta P combinations are shown in Table 45 for Delta P w2 w1 nsubj dependencies while low scoring Delta P w2 w1 nsubj dependencies are shown in Table 46.

Table 45: High Scoring Delta P w2 w1 Nsubj Dependencies

Nsubj Dependency	**Delta P score**
resistance__nn_:_be__vb	0.81
measure__nn_:_have__vb	0.76
belief__nn_:_be__vb	0.73
program__nn_:_have__vb	0.70
fear__nn_:_be__vb	0.63

Table 46: Low Scoring Delta P w2 w1 Nsubj Dependencies

Nsubj Dependency	**Delta P score**
student__nn_:_come__vb	-0.001
thinking__nn_:_be__vb	-0.001
student__nn_:_provide__vb	-0.001

graduate__nn_:_have__vb	-0.001
wave__nn_:_be__vb	-0.001

While an explanation of why these particular units may have a higher likelihood of decreasing grade is not entirely possible from only corpus-based observations, these results along with others (e.g., the positive correlations in Durrant et al., 2019 and Garner et al., 2018, 2019) continue to raise questions about the explanations behind the correlations between these measures of association and student writing scores.

An examination of the negative coefficient t-score nsubj dependency measure in Tables 47-48 also illuminates a similar picture to other negative predictors. Both high and low scoring combinations show that combinations that comprise of two highly frequent words (i.e. nouns such as ‘people’ and verbs such as ‘have’ and ‘be’. This negative result and inspection of high and low scoring combinations makes intuitive sense since these combinations are not particularly indicative of academic writing and are noted to occur across multiple modes of communication and contexts.

Table 47: High Scoring T-score Nsubj Dependencies

Nsubj Dependency	**T-score Score**
analysis__nn_:_be__vb	12.22
participant__nn_:_be__vb	11.46
theory__nn_:_be__vb	11.13
study__nn_:_have__vb	11.08
model__nn_:_be__vb	10.96

Table 48: Low Scoring T-score Nsubj Dependencies

Nsubj Dependency	**T-score Score**
energy__nn_:_ have__vb	-16.53
organization__nn_:_ have__vb	-16.59
school__nn_:_ be__vb	-17.05

time__nn_:_ have__vb	-17.07
government__nn_:_ be__vb	-18.72

Tables 49 and 50 show that a similar picture emerges for Delta P w1 w2 dependencies. The high scoring combinations in Table 49 show some cross-over with high scoring MI combinations. For example, combinations such as ‘shed__vb_:_light__nn’ feature highly on both lists as does ‘bridge__vb_:_gap__nn’, however, there are a number of combinations which have highly ranking Delta P score but much lower MI scores. For example, comparing both the ranking of the MI and Delta P scores, entries such as ‘solve__vb_:_problem__nn and ‘play__vb_:_role__nn’ feature in the top 10 highly ranked combinations for Delta P but when examining the MI scores, these same combinations only feature among the top 40 entries. This raises questions about the relationship between the property of exclusivity that is being measured by the MI measure, and the weigh being placed on word 1 of the combination in the Delta P w1 w2 dobj dependencies. These subtle differences in ranking are worth further examination, possibly with consideration for what cognitively may be being represented for word combinations where word 1 in the combination is more likely to attract word 2. It is also worth considering how and why word 1 being more likely to attract word 2 has a lesser chance of grade increases. These types of quantitative patterns and rankings are one avenue that this thesis flags up as worth further investigation. Table 50 shows that low ranking Delta P combinations comprise of basic or nuclear combinations which do not seemingly indicate use in a particular genre, discipline or topic. These combinations are very much generic in nature.

Table 49: High Scoring Delta Pw1w2 Dobj Dependencies

Dobj Dependency	Delta P w1 w2 Dobj Score
shed__vb_:_light__nn	0.29
bridge__vb_:_gap__nn	0.21
reduce__vb_:_cost__nn	0.22
have__vb_:_sense__nn	0.22
create__vb_:_space__nn	0.21

Table 50: Low Scoring Delta P w1w2 Dobj Dependencies

Dobj Dependency	Delta P w1w2 dobj Score
be__vb_:__ word__nn	-0.033
be__vb_:__ class__nn	-0.034
be__vb_:__ number__nn	-0.034
be__vb_:__ student__nn	-0.034
be__vb_:__ example__nn	-0.034

To sum up, the modelling process has helped draw attention to the types of word combinations that FYC programmes may look to in the future in their drive to make language a more integral part of their instruction.

The other important finding from the statistical modelling was the role played by amod and nsubj diversity measures. Both of these measures were also found to increase the odds of higher grades in the FYC context. In attempting to understand why this diversity leads to higher odds of a grade increase, a qualitative inspection of texts with high amod and nsubj diversity was carried out. I looked at a 20% sample of highly ranked amod and nsubj diversity scores to establish how dependency types were commonly being used to fulfil text functions. These texts had an amod diversity value range from 10.334 -8.471 and a nsubj diversity value range from 10.334 - 6.786. I read through each of the texts and coded examples of the amod and nsubj dependencies in terms of the functions they appear to facilitate. A summary of the broad themes that emerged across the 20% of the texts is shown in examples [32] – [41]:

- Emphasise the importance of the topic generally and historically over time:

[32] In the 1960s people became interested in study of whales and dolphins in the wild and spent **significant time** observing their behavior.

- Generally critique the topic under study:

[33] The next few paragraphs will be describing what happened in Ukraine prior and during the invasion, the failures and importance of energy between the EU and Russia, and why this is more then just a coincidence as many **critics** might **lead** you to believe')

- Make reference to source-based evidence:

[34] Modern studies provide **credible evidences** to support this theory.

- Show support for the efforts of a particular stakeholder:

[35] Government space agencies have rapidly and, for the **most part**, reliably developed technology that improves everyday life.

- Interpret the evidence presented to support a point:

[36] Dr. Davis results show that no matter the regulation provided the **public** will **abuse** of medical marijuana just as they had already abused of the drug in the 1920s'; Lovinger clarifies that, LSB, Cannabinoid receptors generally inhibits neuronal excitability and neurotransmitter (Lovinger 1156) What Dr. Lovinger is saying is that when the human body undergoes a treatment or use of marijuana, the **drug** does not **allow** for the nervous system to continue working in the ways that it should').

- Indicate that there are different sometimes opposing views on the topic:

[37] Additionally, some scientists present **alternative point** of view on the subject.

[38] There is another class of individuals that ignores the factual analysis and presents a **different perspective**.

[39] The researchers have done their research about this subject in afflication with the university of Bergen and the institute of global health and community medicine in Bergen, and come to the assumption that using surrogates in communities such as India, with lower economy and a **different** cultural **view** on the role of females and pregnancy, is something they would strondly advise against.

- Highlight the benefits of taking action:

[40] In most aspects of life, a **common goal** can lead to an unexpected partnership.

[41] The process has the advantage of being in its infancy and the potential to become an even more **accurate procedure.**

The presented functional uses of these dependencies go *some* way to helping students meet the fundamental primary learning objectives of their FYC programme. With reference to the tasks that students need to complete, in the module ENC 1101 task they are expected to synthesise multiple stakeholder perspectives in the form of a literature review and in the module ENC 1102, they are expected to balance the competing views expressed in the ENC 1101 task by attempting to show how stakeholders can work together and reach a compromise (See Chapter Two, Section 2.4). The examples show how the dependency types go *some* way to perhaps facilitating greater task achievement as students navigate the task of setting out the key arguments between stakeholders and then looking at how these stakeholders can reach a viable compromise.

At the same time, these dependencies and their inter-related functions also appear to go some way towards students showing *critical thinking* in that students '*evaluate evidence, recognize and evaluate underlying assumptions, identify and evaluate chains of reasoning, and compose appropriately qualified and developed claims and generalizations*' (CWPA Outcomes Statement, 2014). The Postsecondary Framework for Success (2011) also describes how writers are asked to 'write texts for various audiences and purposes that are informed by research (e.g., to support ideas or positions, to illustrate alternative perspectives, to provide additional contexts).

These results add weight to the established comments attributed to Aull (2015a), Gere (2016) and Matsuda (2006) who all acknowledge that language patterns are not isolated structures that appear randomly but are instead linked to the macro-level writing processes/demands of the assignment and that these language patterns facilitate these processes.

In relating findings from the statistical modelling to the wider body of collocation-grade literature, there is some support for the comments made by Durrant and Brenchley (in press) that phraseological sophistication or indeed the phraseological complexity, when we include the sub-construct of diversity, is not a uniform construct that develops or is evaluated uniformly across educational contexts. As the FYC findings here differ from the CEFR-based work of Paquot (2019), it highlights that what raters tend to focus on and assign high scores to is context sensitive.

Although the present study aligns with previous similar studies in finding support for both the MI and t-score measures, the respective dependencies that have been illuminated are different, and unlike previous studies, this study finds

support for the diversity of both amod and nsubj dependencies as having a greater chance of increasing grade scores. In obtaining these results, there are a number of contextual differences in the study methods and contexts that need to be appreciated and recognised as being able to account for such differences.

One methodological consideration that influences this study's alignment with the literature is clearly the fact that the present study is based on dependency types as opposed to the use of bigrams and trigrams which have been automatically extracted as adjacent pairings in much of the previous literature (with the exception of Paquot, 2018, 2019 who also focuses on dependencies).

In previous university writing studies, there has been a focus on second language proficiency with a focus on evaluating students' proficiency as being adequate enough for university study. This was the case in both Garner et al's (2018, 2019) study of the Korean placement test. Equally, the texts in the present study are long extended pieces of writing which have been extensively drafted and received feedback from both instructors and peers. This type of writing clearly differs from much of the literature which has focused on short timed writing (e.g., Garner's placement tests were timed as were the descriptive essays in Bestgen and Granger (2014). Finally, Paquot's (2018, 2019) work also differs from the present context by focusing on coursework texts written by postgraduate linguistics' students who were evaluated under the CEFR scale. All of these contextual differences highlight the nuanced picture of writing we obtain under such a quantitative analysis.

Further to the writing type, the FYC context also differs in two other ways that may influence why results differ from the literature. First, the FYC context is evaluating a broad range of rhetorical and curriculum goals as opposed to the often-narrower focus on ESL courses and second language courses that have

been the subject of study previously. These types of different foci have been highlighted recently by Lee (2019) who notes that ESL courses cover how to write and place a greater focus on explicit language use while FYC courses are geared to assess multiple rhetorical goals including critical thinking and consideration for audience and at the same time favour implicit and subtle use of language (Lee, 2019, p. 4). Similarly, Ringler, Klebanov and Kaufer (2018) also highlight that the type of argumentative writing that takes place in local FYC contexts differs markedly from the type of argumentative writing that appears in global contexts. In their analysis of FYC writing, they found that it differed in terms of the use of academic language, personal register, assertive language and reasoning. This is also another consideration to bear in mind when drawing parallels with the existing literature on collocation-grade literature.

In relation to Lee's (2019) observations about curriculum goals and different objectives is the second point that the present study found that non-native writers had a higher chance of obtaining a higher grade than their native peers. There are several reasons that could account for this. Although, it could be argued that the reason for this result is that raters could sub-consciously favour non-native writers, it is also a possibility that they are more likely to receive higher grades because they practice their writing more and have had more exposure to writing through preparation courses pre-university and are therefore more in tune with writing for an audience and writing conventions (cf. Lee, 2019). It is also possible that L2 students have a higher chance of receiving a higher grade because of the make-up of the raters. It is a feature of the FYC environment that essays are graded by instructors with a mixed experience level and background in teaching and assessing writing. The make-up of the FYC environment at USF is that there are GTAs with far less experience than other

faculty members and while it may be attested that rater experience is a factor in rater judgements, the result in this FYC context is found to be in conflict with much of the previous literature (e.g., Weigle, 1999) which has found that inexperienced raters are actually more severe in their rating until they receive training to balance out their subjective views. At the same time. rater and writer backgrounds may be another factor which possibly accounts for the FYC result here. However, while previous research has found that L2 writing has been consistently graded lower than L1 writing (Huang & Foote, 2010), the FYC result obtained here seems to indicate that rating behaviour is context sensitive, and only more fully understood when placed in that context. A further exploration of the FYC context is therefore needed, beyond the scope of this thesis, to understand these potential reasons in more nuanced detail.

7.9 Summary

This study has helped highlight the importance of implementing a type of mixed effects modelling which accounts for the random effects structure of the sampling corpus and at the same time preserves the ordinal levels of the outcome/dependent variable. The analysis highlighted how a number of linguistic and non-linguistic fixed effects appeared to have the potential to increase or decrease the odds of essays being awarded higher or lower grades. However, an inspection of the threshold coefficients suggest that these effects are not uniform or operating in the same directions for all grade levels. The final chapter of this study illuminates the contributions to knowledge that the modelling has made and discusses what needs to be considered as a priority for future research.

Chapter Eight: Conclusion

8.1 Main Contributions of the Study

This study set out to firstly explore the relationships that existed between an array of association measures, and secondly the extent that a refined set of these association measures alongside measures of collocation diversity had a relationship with writing quality grade scores in a corpus of FYC student writing. The rationale for such exploration was grounded in three observations that emerged from previous literature:

(1) Previous studies have focused on a narrow range of L2 contexts and little is known about contexts where L1 and L2 writers are taught and assessed under the same programme objectives and assessment criteria. These previous studies also point to the different nature collocation development may take across different assessment contexts.

(2) The literature presents the measures of association that have been used to tap into the sophistication of collocation in a fragmented manner resulting in a need to explore the relationships that exist between the vast array of possible measures that tap into different properties of collocation; and to examine the array of measures which flag up or highlight different collocation types and their respective uses.

(3) Past literature has largely relied on monofactorial statistical methods that have failed to take into account the hierarchical structure of the corpus and how the individual differences that result from individual raters and class structures may introduce and account for variation in the grading process.

Overall, the methods and two studies in this research have contributed to a wider drive in the FYC literature to complement composition instruction with that of language support. In this manner, the study therefore adds an additional voice to this literature. First, it advocates the use of corpus linguistics techniques and second the use of advanced statistical techniques to understand how aspects of collocation may contribute to grade allocation and ultimately facilitate students' task completion and achievement of FYC programme goals. This contribution therefore supports the work of other scholars who have found empirical evidence that language instruction, support and examples should be embedded into FYC student learning and more broadly FYC programme outcomes and guiding frameworks (e.g., Aull, 2015a, 2017, 2019; Eckstein & Ferris, 2018; Perin & Lauterbach, 2018).

The methods employed in this study have contributed to understanding the relationships between collocations and writing quality. In extracting collocational units as dependencies, the study has shed light on how collocations may be extracted by using a relatively lesser known way of extracting collocations in the first and second language learning and assessment literature. In this sense, the manual check of how well these dependency relations were extracted with the Stanford parser (Manning et al., 2014) adds value to our understanding of how well automated linguistic tools perform in a given dataset. Alongside work such as Durrant and Brenchley (in press), Gilquin (2017) and Huang et al (2018), this manual check also allows users to decide on weighing up manual and automated parsing decisions when working with a learner text dataset. Although, the findings of such a check supported a relatively high level of reliability, it is

worth bearing in mind that the number of dependencies captured by this extraction method was still likely to be considerably less than manual identification, when precision and recall scores are examined. Therefore, like the use of other automated tools, researchers need to make a trade-off between speed and accuracy.

Another contribution to knowledge was made in Chapter Six by looking at the relationships between computationally-similar association measures. The analysis choice of the cluster analysis helped tease out similarities and differences between these measures and ultimately inform decisions of choosing distinctly different association measures that are thought to be able to illuminate different properties and uses of collocations. The cluster analysis was particularly useful for setting out and understanding how a large group of selected association measures overlapped in their mathematical similarity and subsequently the extent they have been able to tap into underlying indirect properties of collocation. The analysis was especially useful for understanding differences between coefficient measures which were in fact highly correlated within their cluster. Similarly, the cluster analysis also helped show that hypothesis measures and heuristic measures had particularly high degrees of individuality and therefore low correlation with other measures.

A final contribution to knowledge was made in Chapter Seven through the use of ordinal mixed effects modelling in the form of a cumulative link model. This model has not featured much, if at all, in first and second learner language writing research and in using such a model type that can preserve grade levels, it is hoped that the writing community continues to explore possible avenues for modelling work with this model type in the future. The modelling process contributed to our understanding of the relationship between collocations and

writing quality in a number of ways. The contrast between the fixed and mixed effects models highlighted the importance of incorporating the random effects structure into the model. This indicated that on the whole, the random effects structure was able to contribute more to explaining the variance in grade scoring than the fixed effects of task, language status or the linguistic features. However, an examination of the fixed effects also revealed a number of observations. The modelling process helped reiterate the potentially important role that the MI, t-score, Log-Likelihood Ratio Squared and Delta P may play in evaluating learner language. Although these association measures varied in their significance and direction of relationship with grade, they all, to varying extents, had the effect of increasing or decreasing the odds of a grade increase with their use. To this end, their influence on grade scores seemed to be most prominent at higher grade levels when the threshold coefficients were examined. This seems to indicate that the mixed effects model is more able to predict higher grades off the back of measures of collocation acting as predictors rather than lower grade scores. It also appears to suggest that raters tend to zoom in more on collocation and their respective different properties that are illuminated by the association measures, at higher grade levels. These findings are all areas that future research should consider exploring. The study also adds weight to previous phraseology-grade work in that it also finds that relationships between collocation and writing quality are not homogeneous but instead context specific with other grading factors also playing a role in final grade allocation alongside the use of collocations (e.g., in line with Durrant & Brenchley, in press).

In developing an understanding of the role of dependencies, it is also important to compare the effect of these linguistic predictors vis-à-vis the predictors of task and language status. Importantly, the modelling process illuminated that task was not a significant predictor of grade, however, the language status of the writer did register a weak level of significance (p=0.072). When examining the odds ratio, it is also important to interpret the linguistic predictors as being as having at least a similar effect on grade level as the non-linguistic predictor of language status. In fact, the odds ratio of language status is the same as those of the amod diversity, slightly less than the nsubj diversity and less than the odds of the MI nsubj and MI dobj dependencies. This finding offers an important glimpse into grading practices in the FYC programme.

8.2 Implications for Assessment

As the contributions to knowledge show, the findings have a number of implications for the instruction and assessment taking place on USF's first-year writing programme. These implications relate to future instructional considerations. Chapter Seven's follow up qualitative analysis highlighted how texts with high amod and nsubj diversity contained amod and nsubj dependencies that were used to perform a number of rhetorical functions. The connection between the association measure scores and their diverse use is of key importance to future research. Future research needs to look at how the high and low-scoring association measure combinations are being used to develop a sense of high or low achievement in the FYC tasks. This kind of quantitative-

qualitative examination would help identify how language is being used to facilitate meeting task and wider FYC programme goals. In this regard, such banks of learner language may be invaluable for instructors as complementary aids that supplement the external 'community comments' that already exist within the My Reviewers platform (See Chapter Two, Section 2.5.3).

These authentic language examples would allow FYC instructors to clearly connect their process-oriented rhetorical instruction with examples of learner language that show students how to navigate the tasks and demonstrate ownership of their own texts. This dual approach would ultimately mean that students receive both adequate composition and academic language support; with this language support being directly related to the tasks students are being expected to complete. This is not to advocate that students should be encouraged to simply repeat the language they are exposed to; instead they should be encouraged to notice the wide array of language choices they have available to them in conveying their intended meaning. They should also be encouraged to make choices that are befitting of the tasks being completed.

8.3 Limitations and Directions for Future Work

It is important to interpret the study's findings with a degree of caution and awareness of the caveats that apply to the research design. A first acknowledgement must be made in that the measurement of sophistication in the form of association measures is limited by two study design decisions. The first being the choice to measure sophistication using only association measures. There is a clear need to treat the construct in the future in a more multidimensional way that includes more than one measure type. This may

include looking at lists of academic collocations (e.g., in a similar manner to Paquot's (2019) approach). The second study design decision that shapes the view of sophistication obtained here is that of the choice of reference corpus. While the choice of reference corpus was appropriate for learner writing, this limited the view of sophistication by only looking at combinations present in another specialised corpus of writing. A very different picture of sophistication may be presented when the reference corpus is general in nature (e.g., use of the BNC or COCA).

Further to this, the study sample was restricted by the complete responses from learners with regards to gaining insight into their demographic profiles. This information was gathered via a voluntary student survey and this means that the inferences we can draw from these learner variables are limited to having enough complete information. The reliance on such a voluntary system means that future modelling work will need to be carried out with this issue of patchy data across variables in mind. Although mixed-effects modelling has been promoted as a statistical technique that can handle missing data, in the case of the FYC database, there are, depending on researcher interest, cases where absent data is the norm, and careful data exploration will be needed in future work to determine exactly the amount of missing data such modelling is able to handle. As a field, this is an issue that learner corpus research is only starting to explore and as a community it is an endeavour that should be pursued for the foreseeable future in our work (e.g., in line with recommendations made by Gries, 2015, 2018).

A further limitation and avenue for future research relates to the exclusion of topic as a viable variable in the study. In the FYC programme, students are permitted to choose their own topics meaning that inferences depend on enough students choosing similar or the same topics. Although topic has long been recognised as a viable source of individual variance (and therefore treatable as random variance), grouping these topics together was not possible because of the vast array of different topics that students chose. However, in the future the inclusion of topic as a random variable may be possible in a larger or different FYC sampling frame.

There are a number of other directions that future research should also be encouraged to pursue in light of the findings in this study. A key direction that researchers should be encouraged to explore is the relationships between context specific measures of association. While the present study opted to focus on computationally-simple measures of association, there is clear scope for carrying out further relationship-based work that looks at how mathematically complex context measures may tap into different aspects of context and collocation properties (e.g., see Gries & Durrant's (2020) support for KL divergence as a viable candidate of study).

With respect to the FYC context, the present study represents one of the first collocation-grade studies in the FYC literature and future replications of this work across other FYC contexts would strengthen claims that instruction and assessment on these programmes would benefit from being informed by (a) corpus linguistics techniques and (b) EAP pedagogic methodologies that home in on language as a central component of composing texts.

Embedded within this continuing work is also the recognition that the value of collocations needs to be ascertained when placed along with measures of single word and grammatical structure measures. This kind of holistic analysis may further validate our understanding of collocation by illuminating their strength or weakness as predictors vis-à-vis single word and grammatical structure choices (e.g., Paquot, 2019). In our own wider review of writing quality we have noted that measures (and the knowledge/production they claim to tap into) of syntax, vocabulary, phraseology and cohesion are all alternative perspectives on language and ultimately its importance to writing (Durrant, Brenchley & McCallum, 2021). Their focus depends on the researcher's paradigmatic stance on writing proficiency however in the case of collocation it is worthwhile to see how our view through this particular lens competes with or complements the understanding of writing quality that we obtain from looking at single words/structures.

A penultimate direction that quantitative work like this needs to take is the acknowledgement that these exploratory patterns are obtained indirectly via corpus data and their interpretation is limited if only the quantitative data is used to look at the rationale for such patterns. To this end, this kind of corpus research should be viewed as a starting point for further qualitative exploration of the construct of writing proficiency. The patterns and language examples may be further used in psycholinguistics and interview-based research to tap into why these language examples and patterns of positive or negative relationship with writing quality scores may occur. This seems an especially important step with regards to association measures where simply examining high and low scoring

ranking patterns has limited value in our interpretation of these relationships. This kind of further research also seems to offer an important way forward in trying to understand the continued relationship that we find (e.g., also found in Durrant et al., 2019; Garner et al., 2019) between writing quality scores and the fairly recent examination of the Delta P directional association measure. Since researchers have noted the clear psycholinguistic property of Delta P (e.g., see comments made in Chapter Four in the work of Schneider, 2018), follow up research that is qualitative in nature seems to offer promise in helping us explore this relationship further.

A final direction for this work relates to more broadly how these linguistic predictors have a relationship to other contextual predictors that represent the grading process more holistically. The modelling process indicates that L2 writers have a higher chance of being at or beyond a particular grade level and to some extent this quantitative-heavy modelling also supports the qualitative picture built up in previous grading bias studies (e.g., Huang & Foote, 2010; Brown, 1991). Similar to the suggestions already made around a qualitative focus to complement these findings, there is a need to investigate why L2 writers have a higher chance of an increased grade. A qualitative focus on this rationale may help illuminate potential trajectories of rater bias in the FYC programme.

On the whole, the modelling process has illuminated a number of statistical patterns that should be further investigated through qualitative means that ultimately provide a fine-grained understanding of the relationship between collocations, dependencies and learner and contextual FYC programme variables. This is a highly viable direction that future FYC work should be encouraged to take.

Appendix A

FYC Rubrics

Table 51: ENC 1101 Project 3 Joining the Conversation Rubric

Criteria	Emerging (0-2)	Developing (3-5)	Mastering (6-8)
Analysis 25%	• Assignment requirements not met • Thesis absent or minimally presents arguable claim • Little or no connection between thesis and claims presented in essay • Little or no development of supporting points relative to arguable claim	• Assignment requirements partially met • Thesis partially presents arguable claim • Partial connection between thesis and claims presented in essay • Partial development of supporting points relative to arguable claim	• Assignment requirements adequately met • Thesis presents arguable claim • Adequate connection between thesis and claims presented in essay • Adequate development of supporting points relative to arguable claim
Evidence 25%	• Arguable claims minimally supported by appropriate and credible sources • Supporting details minimally relevant to arguable claims • Source material not properly integrated • Quotes, paraphrases, and summaries improperly cited • Little distinction between writer's voice and source's ideas	• Arguable claims partially supported by appropriate and credible sources • Supporting details partially relevant to arguable claims • Source material inconsistently integrated • Quotes, paraphrases, and summaries inconsistently cited • Some distinction between writer's voice and source's ideas	• Arguable claims supported by appropriate and credible sources • Supporting details relevant to arguable claims • Source material consistently integrated • Quotes, paraphrases, and summaries properly cited • Adequate distinction between writer's voice and source's ideas

Organization 20%	• Opening presents minimal background information on topic and problem • Topic sentences absent or minimally relevant to thesis and paragraph's content • Transitions absent or infrequently used • Supporting points flow illogically • Conclusion absent or irrelevant to thesis and arguable claims	• Opening presents partial background information on topic and problem • Topic sentences somewhat relevant to thesis and paragraph's content • Transitions inconsistently used • Supporting points flow somewhat logically • Conclusion somewhat relevant to thesis and arguable claims	• Opening presents adequate background information on topic and problem • Topic sentences consistently relevant to thesis and paragraph's content • Transitions consistently used • Supporting points flow logically • Conclusion relevant to thesis and arguable claims
Format 15%	• Document design for header, heading, line spacing, margins, and font style minimally compliant with MLA style conventions • Little attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display incomplete source information	• Document design for header, heading, line spacing, margins, and font style partially compliant with MLA style conventions • Inconsistent attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display partially complete source information	• Document design for header, heading, line spacing, margins, and font style compliant with MLA style conventions • Consistent attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display complete source information
Style 15%	• Significant problems with sentence construction, diction, and word choice • Frequent grammar and punctuation errors • Frequent proofreading errors • Inconsistent point of view • Language significantly interferes with communication of ideas	• Some problems with sentence construction, diction, and word choice • Some grammar and punctuation errors • Some proofreading errors • Somewhat consistent point of view • Language occasionally interferes with communication of ideas	• Few or no problems with sentence construction, diction, and word choice • Few or no grammar and punctuation errors • Few or no proofreading errors • Consistent point of view • Language facilitates communication of ideas

Table 52: ENC 1102 Project 1 Finding Common Ground Rubric

Criteria	Emerging (0-2)	Developing (3-5)	Mastering (6-8)
Analysis 25%	• Assignment requirements not met • Ideas and assertions minimally conform to Rogerian style of argument • Thesis absent or minimally represents potential for compromise between stakeholders • Individual stakeholder positions minimally represented • Common ground minimally established • Ideas for compromise minimally developed	• Assignment requirements partially met • Ideas and assertions partially conform to Rogerian style of argument • Thesis partially represents potential for compromise between stakeholders • Individual stakeholder positions partially represented • Common ground partially established • Ideas for compromise partially developed	• Assignment requirements met • Ideas and assertions adequately conform to Rogerian style of argument • Thesis adequately represents potential for compromise between stakeholders • Individual stakeholder positions adequately represented • Common ground adequately established • Ideas for compromise adequately developed
Evidence 25%	• Source research minimums not met • Arguable claims minimally supported by appropriate and credible sources • Supporting details irrelevant to stakeholders' positions and claims • Source material improperly integrated • Quotes, paraphrases, and summaries improperly cited • Little distinction between writer's voice and source's ideas	• Source research minimums partially met • Arguable claims partially supported by appropriate and credible sources • Supporting details somewhat relevant to stakeholders' positions and claims • Source material inconsistently integrated • Quotes, paraphrases, and summaries inconsistently cited • Some distinction between writer's voice and source's ideas	• Source research minimums adequately met • Arguable claims adequately supported by appropriate and credible sources • Supporting details relevant to stakeholders' positions and claims • Source material properly integrated • Quotes, paraphrases, and summaries properly cited • Adequate distinction between writer's voice and source's ideas

Organization 20%	• Opening minimally introduces topic and two stakeholder positions • Topic sentences absent or minimally relevant to thesis and paragraph's content • Transitions absent or minimally used • Supporting points flow illogically • Conclusion absent or minimally relevant to thesis and arguable claims	• Opening partially introduces topic and two stakeholder positions • Topic sentences inconsistently relevant to thesis and paragraph's content • Transitions inconsistently used • Supporting points flow somewhat logically • Conclusion somewhat relevant to thesis and arguable claims	• Opening adequately introduces topic and two stakeholder positions • Topic sentences consistently relevant to thesis and paragraph's content • Transitions adequately used • Supporting points flow logically • Conclusion relevant to thesis and arguable claims
Format 15%	• Document design for header, heading, line spacing, margins, and font style minimally compliant with MLA style conventions • Little attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display incomplete source information	• Document design for header, heading, line spacing, margins, and font style partially compliant with MLA style conventions • Inconsistent attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display partially complete source information	• Document design for header, heading, line spacing, margins, and font style compliant with MLA style conventions • Consistent attention to MLA formatting of source citations, including hanging indent, punctuation, capitalization, and italics use • Source citations display complete source information
Style 15%	• Significant problems with sentence construction, diction, and word choice • Frequent grammar and punctuation errors • Frequent proofreading errors • Inconsistent point of view	• Some problems with sentence construction, diction, and word choice • Some grammar and punctuation errors • Some proofreading errors • Somewhat consistent point of view	• Few or no problems with sentence construction, diction, and word choice • Few or no grammar and punctuation errors • Few or no proofreading errors • Consistent point of view

	• Language interferes with communication of ideas	• Language occasionally interferes with communication of ideas	• Language facilitates communication of ideas

Appendix B

Lexical Studies

Table 53: Measures of Diversity

Sub-category	Measure	Notes on Operationalization	Study	Findings
Types	All types	Types per essay	Banerjee, Franceschina & Smith 2007	Positive correlation with proficiency (Task 1: r=.44; Task 2: r=.53)
			Vidakovic & Barker 2010*	Increases across proficiency levels
			Ruegg, Fritz & Holland 2011	Individual correlations not provided. In overall regression with other predictors, does not make a significant contribution to prediction of lexis score.
			Kim 2014	Significant increase with proficiency level.
			Douglas 2015	Significant increase for proficiency levels (r=.77)
			Treffers-Daller, Parslow and Williams 2018	Significant positive correlation with writing score (r=.47).
	Lexical types	Total types of nouns, verbs, adjectives and adverbs	Ferris 1994	Increases with essay grade
			Grant & Ginther 2000*	Increases with essay grade
			Santos, Verspoor & Nerbonne 2013	Significant relationship with essay quality
			Kim 2014	Non-significant increase with proficiency level.
		Separate counts for each part of speech	Ferris 1994	Increases with essay grade
			Grant & Ginther 2000*	Increases with essay grade
			Espada-Gustilo 2011	Significant linear increase with proficiency level.

			Kim 2014	Non-significant correlation with proficiency level.
TTR	TTR all words		Cumming & Mellow 1996	No significant difference between advanced and intermediate students.
			Cumming, Kantor, Baba, Erdosy, Eouanzoui & James 2005	Correlation with essay quality but small effect size (np^2=.08).
			Banerjee, Franceschina & Smith 2007	No significant relationship with proficiency.
			Espada-Gustilo 2011	Significant increase with proficiency level.
			Kim 2014	Negative correlation with proficiency level.
			Wang 2014	No significant relationship with quality
			Treffers- Daller, Parslow & Williams 2018	Significant positive correlation with writing score (r=.46).
	Lexical TTR	TTR for nouns, verbs, adjective and adverbs	Nihalani 1981	No significant difference between proficiency levels.
			Engber 1995	Significant positive correlation with essay grade when errors were included (r=.45) and when errors were excluded (r=.57).
	TTR per segment	200 words	Becker 2010	Significant increase with proficiency: Level 1 = Level 2 < Level 3
		50-word segments	Grant & Ginther 2000*	Increases with essay grade
			Vidakovic & Barker 2010*	Increases across proficiency level.

Modified measures	Corrected TTR	Type ÷ (2 * tokens)^2	Arthur 1979 (study 1)	
			Arthur 1979 (study 2)	No significant correlation with proficiency ratings awarded by teachers.
			Kim 2014	Significant increase with proficiency level.
	Root TTR (Guiraud's Index)	# types ÷ √tokens	Verspoor, Schmid & Xu 2012	Significant correlation with proficiency (r =.71)
			Bulté & Housen 2014	Significant increase with quality (r=.52)
			Kim 2014	Significant increase with proficiency level.
			Verspoor, Lowie, Chan & Vahtrick, 2017	Non-significant correlation with essay grade (r=.17)
			Treffers- Daller, Parslow & Williams 2018	Significant positive correlation with writing score (r=.47).
	MAAS index		Bestgen 2017	Significant correlation with quality for both FCE (r=.23) and ICLE (r=.51) texts.
	MTLD	♯ unique words ÷ ♯words with resetting once a pre-set ratio is achieved for total words to that	Aryadoust 2016	Inversely related to essay grades (regression weights -.196, -.053, -.094 at three different testing points)

	point [cf. McCarthy & Jarvis, 2010]	Bestgen 2017	Significant positive correlation with quality for both FCE (*r*=.19) and ICLE (*r*=.44) texts.
		Treffers - Daller, Parslow & Williams 2018	Significant positive correlation with writing score (r=.34).
D		Jarvis 2002	Significant correlation with quality for whole texts (r_s= .34) and content words only (r_s= .36).
		Yu 2010	Significant correlate of quality for writing (r =.29). Regressions show relationship to vary by gender of writer (much higher for males than females) and candidate's purpose in taking the test (much higher for those seeking college admission than those seeking professional certification).
		Crossley & McNamara 2012	Significant positive correlation with quality (r=.43).
		Guo, Crossley & McNamara 2013	Significant positive correlation with score in independent essay (r=.42)
		Bulté & Housen 2014	No significant relationship with quality.
		Wang 2014	No significant relationship with quality
		Aryadoust 2016	Inversely related to essay grades (regression weights -.235, -.089, -.067 at three different testing points)

			Krzeminska-Adamek 2016	No significant correlation with quality ratings at time 1 (r=.03) or time 2 (r=.00).
			Qin & Uccelli 2016	Positive correlation with quality for argumentative (r=.48) and narrative texts (r=.44).
			Yoon 2017	Significant increase across levels (η_p^2=.04).
			Treffers- Daller, Parslow & Williams 2018	Significant positive correlation with writing score (r=.31).
	HD-D	NB. his measure shows extremely high correlation with D (r=.97) in McCarthy & Jarvis, (2010) and in Treffers-Daller, Parslow & Williams (2018) (r=.93).	Bestgen 2017	Significant correlation with quality for both FCE (*r*=.16) and ICLE (*r*=.46) texts.
			Treffers- Daller, Parslow & Williams 2018	Significant positive correlation with writing scores (r=.30).
	Uber-Index	logtokens² ÷ logtokens minus the log types	Jarvis 2002	Significant positive correlation with quality (r_s=.29) for whole texts; and for content words only (r_s= .43).
Pooled measures of diversity	Pooled measure of diversity	Pooled measure of diversity consists of: TTR, CTTR, Root TTR, Bilog TTR and MTLD	Vajjala 2018	Significant positive correlation with proficiency level for TOEFL 11 essays (r=.67) and FCE essays (r=.29).

Table 54: Frequency: List-based Measures

Measure	Notes on Operationalization	Study	Findings
% tokens in BNC 1K	Lexical Frequency Profile (Laufer & Nation, 1993)	Laufer & Nation 1995	Significant decrease with increasing proficiency.
	Range (Nation & Heatley, 1996)	Gregori-Signes & Clave-Arroitia 2015*	Lower in high-proficiency (4th year) than low-proficiency (1st year) texts.
% tokens in GSL 1K	VocabProfile (Cobb, 2019)	Vidakovic & Barker 2010*	No clear trend across proficiency levels
		Yousofi & Bahramlou 2014	Non-significant negative correlation with writing quality (r=-.06).
		Biber & Gray 2013	Slight reduction with increased proficiency in independent task (level 1=86%; level 4 = 82%). No trend in integrated task (level 1=80%; level 4 = 82%). No inferential analysis
% tokens in BNC 2K	Lexical Frequency Profile (Laufer & Nation, 1993)	Laufer & Nation 1995	No significant difference in with increasing proficiency
	Range (Nation & Heatley, 1996)	Gregori-Signes & Clave-Arroitia 2015*	Higher in high-proficiency (4th year) than low-proficiency (1st year) texts.
% tokens in GSL 2K	VocabProfile (Cobb, 2019)	Vidakovic & Barker 2010*	Decreases across proficiency levels.
		Yousofi & Bahramlou 2014	Significant negative correlation with writing quality (r=-.27).

% tokens in combined BNC+COCA 1+2K lists	VocabProfile (Cobb, 2019)	Douglas 2015	Significant negative correlation with proficiency (*r*=-.73)
% tokens not in BNC 1+2K	Lexical Frequency Profile (Laufer & Nation, 1993)	Lemmouh 2008	Significantly higher in distinctions vs. pass-grade literature essays. No significant difference between pass and distinction grade linguistics essays.
		Ruegg, Fritz & Holland 2011	Individual correlations not provided. In overall regression with other predictors, does not make a significant contribution to prediction of lexis score.
		Bestgen 2017	Significant positive correlation with quality for both FCE (*r*=.13) and ICLE (*r*=.28) texts.
	Range (Nation & Heatley, 1996)	Banerjee, Franceschina & Smith 2007	Significant decrease with increasing proficiency for IELTS writing task 2 but not task 1.
% tokens not in combined BNC+COCA 1-3K lists	VocabProfile (Cobb, 2019)	Krzeminska-Adamek 2016	Significant positive correlation with proficiency for second, but not first, set of texts (r=.36).
% tokens in combined BNC+COCA 3-10K lists	VocabProfile (Cobb, 2019)	Douglas 2015	Significant positive correlation with proficiency (*r*=.73)
% tokens in combined BNC+COCA 11-25K lists + off-list words	VocabProfile (Cobb, 2019)	Douglas 2015	Significant positive correlation with proficiency (*r*=.28)
% tokens not in BNC1+2K or University Word List	Lexical Frequency Profile (Laufer & Nation, 1993)	Laufer & Nation 1995	Significant increase with increasing proficiency.
		Bestgen 2017	Significant correlation with quality for both FCE (*r*=.11) and ICLE (*r*=.48)
% tokens not in GSL1+2K or Academic Word List	Range (Nation & Heatley, 1994)	Gregori-Signes & Clave-Arroitia 2015*	Higher in high-proficiency (4th year) than low-proficiency (1st year) texts.

	VocabProfile (Cobb, 2019)	Yousofi & Bahramlou 2014	Significant positive correlation with writing quality (r=.18).
LFP index	Words on the LFP and those absent from the LFP lists are allocated scores by researchers. Words on word list 1 = value of 1, words on word list 2 = value of 2. Words not on the wordlists were assigned a value of 5. Type percentages for each list were then multiplied by the value assigned to that list and summed to give a lexical frequency profile per text.	Ruegg, Fritz & Holland 2011	Individual correlations not provided. In overall regression with other predictors, does not make a significant contribution to prediction of lexis score.
# frequency bands accessed in the task	# frequency bands accessed in producing the text	Douglas 2015	Significant positive correlation with proficiency level (r^2=.58).
Lexical stretch	Lowest frequency band required to achieve 98% coverage of text. VocabProfile (Cobb, 2019)	Douglas 2015	Significant positive correlation with proficiency level (r=.57).
	Lowest frequency band used in text. VocabProfile (Cobb, 2019)	Douglas 2015	Significant positive correlation with proficiency level (r=.45).
S	The frequency level at which text coverage reaches 100% (Kojima & Yamashita 2014)	Bestgen 2017	Significant positive correlation with quality for both FCE (*r*=.12) and ICLE (*r*=.35) texts.
Word difficulty within P_lex tool	Counts of 'hard' words. Hard words are operationalised by counting all words not found on Nation's (1984) first 1,000 words.	Meara and Bell 2001	Significant predictor of quality when text length exceeded 120 words (r=.57)
		Moreno Espinosa 2005	Significant predictor of quality among low-rated texts (r=.18) but not higher rated texts (r= -.07)
Advanced Guiraud Index		Bulté & Housen 2014	No significant correlation with quality.

Verb sophistication	Sophisticated verb variation 1 (from Lu's Lexical Complexity Analyser, 2012) # verb types not in the most frequent 2K in BNC / # verbs	Kim 2014	Significant positive correlation with proficiency levels (η^2=.31)
	Sophisticated verb variation 2 (from Lu's Lexical Complexity Analyser, 2012) # square(verb types not in the most frequent 2K in BNC) / # lexical words).	Kim 2014	Increases with proficiency level but does not appear as a significant correlate with proficiency level
	Corrected sophisticated verb variation (from Lu's Lexical Complexity Analyser, 2012) # verb types not in the most frequent 2K in BNC ² / # verbs)	Kim 2014	Increases with proficiency level but does not appear as a significant correlate with proficiency level
	Verb tokens outside BNC 1+2K. Uses lextutor (Cobb, 2019) adaptation of Lexical Frequency Profile (Cobb, 2019). Includes only verbs from Paquot's Academic Keyword List (Paquot, 2010)		
Internal frequency bands	Percentage of items in frequency bands based on the study corpus	Verspoor, Schmid & Xu 2012	No clear relationship with proficiency.

Table 55: Frequency Measures: Mean Frequencies

Measure	Notes on Operationalization	Study	Findings
CELEX log frequencies (all words)	Coh-Metrix	Yoon 2017	No significant difference across proficiency levels
CELEX frequencies (content words)	Coh-Metrix	Crossley & McNamara 2012	Significant negative correlation with essay grade (r=-.34)
		Green 2012	No significant differences between L2 low and high proficiency levels.
		Guo, Crossley & McNamara 2013	Significant negative correlation with score in integrated essay (r=-.44) and independent essay (r=-.30)
CELEX written frequencies (content words)	Coh-Metrix	Crossley, Salsbury & McNamara 2011	Significant decrease with increasing level (np2=.29).
SUBTLEXus frequencies content words	TAALES	Kyle & Crossley 2016	Significant negative correlation with quality for independent TOEFL essays (r=-.40) but no correlation with integrated TOEFL essays.
Kucera-Francis frequencies (content words log)	TAALES	Kyle & Crossley 2016	Significant negative correlation with quality in independent essays (r= -.36)
Kucera-Francis frequencies (function words log)	TAALES	Kyle & Crossley 2016	Significant positive correlation in independent TOEFL essays (r= .27)
Kucera-Francis frequencies (# categories all words)	TAALES	Kyle & Crossley 2016	Significant negative correlation with quality in integrated essays (r = - .18)
Thorndike-Lorge frequencies (all words, log)	TAALES	Kyle & Crossley 2016	Significant negative correlation with quality in integrated essays (r=-.17)
Spoken BNC frequencies (content words)	TAALES	Kyle & Crossley 2016	Significant negative correlation with quality scores for independent (r= - .34)
Spoken BNC frequencies (function words)	TAALES	Kyle & Crossley 2016	Significant positive correlation with quality in independent essays (r=.11)
COCA academic log frequency	TAALES	Kim & Crossley 2018	Significant negative correlation with independent essay scores (r =-.22) Non-significant negative correlation with source-based essay scores (r= -.09)

Table 56: Register-based Wordlists

Measure	Notes on Operationalization	Study	Findings
% tokens from the university word list	Lexical Frequency Profile (Laufer & Nation, 1993)	Laufer & Nation 1995	Significant increase across proficiency groups
Tokens & types from university word list		Gregori-Signes & Clave-Arroitia 2015*	Increases over proficiency level
% tokens from Academic Word List (Coxhead, 2000)	Range (Nation & Heatley, 1996)	Banerjee, Franceschina & Smith 2007	As proficiency level increases IELTS test takers use more less frequent words from the AWL.
	VocabProfile (Cobb, 2019)	Vidakovic & Barker 2010*	Increases with proficiency level.
		Yousofi & Bahramlou 2014	Significant positive correlation with writing quality (r=.26).
		Verspoor, Lowie, Ping-Chan & Vatrick 2017 (study 1)	Significant positive correlation with essay grade (r=.69).
		Biber & Gray 2013	No change across proficiency levels.

Table 57: Checks with Native Corpora (Range Measures)

Measure	Notes on Operationalization	Study	Findings
Word range (COCA academic)	TAALES	Kim & Crossley 2018	Non-significant negative correlation with source-based essays (r =-.08). Significant negative correlation with independent essay scores (r = -.26)
BNC written range all words	TAALES	Kyle & Crossley 2016	Significant correlation with quality for independent writing task (r=-.41)
SUBTLEXus range (content words)	TAALES	Kyle & Crossley 2016	Correlation with quality score for TOEFL independent essays (r=-40).
Kucera-Francis number of categories (all words)	TAALES	Kyle & Crossley 2016	Significant correlation with quality in integrated essays (r=-.18).

Table 58: Measures of Density

Measure	Operationalization	Study	Findings
Lexical density	adjectives+nouns+verb/all words	Nihalani 1981	No significant differences between proficiency levels
		Engber 1995	No significant relationship with writing quality
		Becker 2010	No significant relationship with proficiency
		Vidakovic & Barker 2010	No clear differences between CEFR levels (No inferential analysis)
		Kim 2014	Significant correlation with proficiency level (η^2=.21).
		Gregori-Signes & Clave-Arroitia 2015* (NB. We are assuming that this is the measure used, though the description in the article appears to confuse density with diversity; specifically with type-token ratio)	Decreases with increasing proficiency
Weighted lexical density	Content words/functions words, with high-frequency items given half the weight of low-frequency items	Banerjee, Franceschina & Smith 2007	Increases across proficiency levels [NB. Though the authors make this claim, it's not clear that the data really support it. Significant differences across band levels go down as well as up]
Content words per clause		Lee 1992	No significant relationship with writing quality

Appendix C

Measures of Phraseology

Table 59: Internal Measures of Collocation

Sub-category	Measure	Notes on Operationalization	Study	Findings: Proficiency
Collocations	Collocations with 5 high frequency verbs: get, give, have, make and take.	Collocate = context words appearing within R3 of verb in more than 10 texts and frequency > 5/100K words	Biber & Gray 2013*	Collocation use differs with task: more collocations used in independent than integrated tasks. No clear link with task score
	Root TTR Based on all combinations meeting POS requirements	types adjective noun collocations ÷ √tokens adjective Based on all combinations meeting POS requirements	Paquot 2019	No significant correlation with CEFR levels.
		types adverbial modifier collocations ÷ √tokens adverbial modifier combinations		No significant correlation with CEFR levels.
		types verb + direct object collocations ÷ √tokens verb + direct object combinations		No significant correlation with CEFR levels.

Table 60: Frequency-based Measures

Sub-category	Measure	Notes on Operationalization	Study	Findings
% attested combinations	Proportion of attested bigrams (written BNC)	TAALES index (proportion of bigrams in learner text that are amongst the 'N' most frequent bigrams in the BNC)	Kyle & Crossley 2016	Significant positive correlation with quality for independent TOEFL essays (r=.15)
	Proportion of attested trigrams (written BNC)	TAALES index (proportion of bigrams in learner text that are amongst the 'N' most frequent bigrams in the BNC)	Kyle & Crossley 2016	Significant positive correlation with quality for independent TOEFL essays (r=.22)
	Proportion of attested trigrams (COCA academic)	TAALES index (proportion of bigrams in learner text that are amongst the 'N' most frequent bigrams in the COCA academic)	Garner, Crossley & Kyle 2019	Significant positive correlation with writing proficiency CEFR levels (r=.34).
	Proportion of absent bigrams (COCA: types)	Proportion of bigram types that are not present in COCA	Bestgen & Granger 2014	Significant negative correlation with overall quality (r = -. 28), language (r=-.37) and vocabulary (r=-.16) scores.
			Bestgen 2017	Significant negative correlation with quality (r = -21) for FCE texts.
				Non-significant positive correlation with quality (r =.11) for ICLE texts.

	Proportion of absent bigrams (COCA: tokens)	Proportion of bigram tokens that are not present in COCA	Bestgen & Granger 2014	Significant negative correlation with overall quality (r = -. 27), language (r=-.36) and vocabulary (r=-.15) scores.
	Proportion of bigrams and trigrams	Proportion of bigram and trigram proportions as a PCA component	Kim, Crossley & Kyle 2018	Significant positive correlation with writing proficiency scores (r=.19) and lexical proficiency scores (r=.24).
	Proportion of combinations attested in source materials	Prompt and non-prompt-based lexical bundles	Staples, Egbert, Biber & McClair 2013	Lower proficiency levels used more prompt-based bundles (η^2 = .04).
				higher proficiency levels used more non-prompt bundles (η^2 = .02).
			Appel & Wood 2016*	Lower proficiency levels used more prompt-based bundles.
		3-word combinations found in source materials	Cumming, Kantor, Baba, Erdosy, Eouanzoui & James 2005	Fewer word strings used as proficiency level increased for reading-writing integrated tasks. Proficiency level used fewest strings on the politics and cinema tasks overall.
				More word strings used as proficiency level increases with listening-writing integrated tasks.
				On the cinema task, proficiency level 3 used more

				word strings than proficiency levels 4 and 5. At proficiency level 4, more strings were used in the Politics reading-writing task than the cinema reading-writing task but counts were roughly equal in both listening-writing tasks. Proficiency level 3 texts contained many phrases for both reading-writing tasks but few for listening-writing tasks. On the listening-writing tasks, levels 4 and 5 used more strings than level 3.
		Number of collocations copied from source. Specifically, collocations with 5 high frequency verbs: get, give, have, make and take. Collocate = context words appearing within R3 of verb in more than 10 texts and frequency > 5/100K words	Biber & Gray 2013	More in integrated than independent task No clear link with task score (Descriptive statistics only)
% combinations below frequency threshold	Below threshold adjective modifier noun	Infrequent combination types in BNC reference corpus (appearing less than 5 times)	Granger & Bestgen 2014	Significantly more used at advanced proficiency level for types (cohen's d=.46) and tokens (cohen's d=.49)
		Infrequent combination types in the L2RC reference corpus (appearing less than 5 times)	Paquot 2018	No change between B2-C2 CEFR levels.

	Below threshold noun + noun	Infrequent combination types in BNC reference corpus (appearing less than 5 times)	Granger & Bestgen 2014	More used at intermediate proficiency level but not statistically significant for types or tokens
	Below threshold adverb pre-modifying adjective	Infrequent combinations types in BNC reference corpus (appearing less than 5 times)	Granger & Bestgen 2014	Significantly used more at advanced proficiency level for types (cohen's d=.60) and tokens (cohen's d=.62)
		Infrequent combination types in the L2RC reference corpus (appearing less than 5 times)	Paquot 2018	Non-significant increase between B2-C2 CEFR levels.
	Below threshold verb + direct object types	Infrequent combination types in the L2RC reference corpus (appearing less than 5 times)	Paquot 2018	Non-significant decrease between B2 – C2 CEFR levels.
	Below threshold: adjective + noun & noun + noun combinations	Infrequent combination types in BNC (less than 5 occurrences in reference corpus) therefore cannot be assigned an MI score	Granger & Bestgen 2014	More beyond threshold combinations used by advanced learners than intermediate learners for types (cohen's d = .29) and tokens (cohen's d = .34)
Mean frequency	Normed mean log frequency of bigrams (written BNC)	TAALES index	Kyle & Crossley 2016	Significant positive correlation with quality for independent TOEFL essays (r=.11)
				Significant negative correlation with quality for integrated TOEFL essays (r = -12)
	Normed mean log frequency of academic trigrams	TAALES index	Garner, Crossley & Kyle 2019	Significant positive correlation with writing proficiency CEFR levels (r=.12).

	Bigram frequency and range as a component in PCA analysis.	Bigram frequency and range calculated as a component in PCA analysis by calculating range and frequencies from COCA sub-corpora, the BNC and HAL. Component also includes 2 x MI measures	Kim, Crossley & Kyle 2018	Significant positive correlation with writing proficiency scores (r=.21) but not lexical proficiency scores (r=.15)

Table 61: Range-based Measures

Sub-category	Measure	Notes on Operationalization	Study	Findings
Bigram range	Bigram range (academic)	TAALES index	Garner, Crossley & Kyle 2019	Significant positive correlation with CEFR writing proficiency levels (r=.23)
Trigram range	Trigram log range (academic)	TAALES index	Garner, Crossley & Kyle 2019	Significant positive correlation with CEFR writing proficiency levels (r=.13)

Table 62: Formula Lists

Sub-category	Measure	Notes on Operationalization	Study	Findings
Cross-checks with lists of formulas	Academic Collocations List	Presence or absence on this list for adjective and noun collocations	Paquot 2019	No significant increase between proficiency levels
		Presence or absence on this list for adverbial modifier and verb collocations		Non -significant increase between proficiency levels
		Presence or absence on this list for verb and direct object collocations		No significant increase between proficiency levels
Academic formulaic language components	Academic formulaic language component in PCA analysis	Component is made up of cross-checking the Academic Formula List (all formulas and core formulas) as well as calculating the COCA academic range logarithm and the COCA academic trigram unigram to bigram association strength (MI^2)	Kim, Crossley & Kyle 2018	Significant positive correlation with writing proficiency scores (r=.19) and lexical proficiency scores (r=.17)

Table 63: Measures of Association

Sub-category	Measure	Notes on Operationalization	Study	Findings
MI threshold measures	% high-MI combinations (MI ≥ 7)	all bigrams	Granger & Bestgen 2014	Significantly more high MI collocation types and tokens used at advanced than intermediate proficiency level (cohen's d= .89 for types; d=.84 for tokens)
		premodifier+noun bigrams	Granger & Bestgen 2014	Significantly more used at advanced than intermediate proficiency level for types (cohen's d=.60) and tokens (cohen's d=.51)
		adjective+noun bigrams	Granger & Bestgen 2014	Significantly more used at advanced than intermediate proficiency level for types (cohen's d=.47) and tokens (cohen's d=.45).
		adjective+noun combinations with modifier dependency	Paquot 2018	Types only: Non-significant increase between B2-C2 CEFR levels.
		noun+noun bigrams	Granger & Bestgen 2014	Significantly more used at advanced than intermediate proficiency level for types (cohen's d=.48) and tokens (cohen's d=.43)
		adverb+adjective bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens.
		adverb+adjective & adverb-verb combinations with modifier dependency	Paquot 2018	Types only: No change across B2-C2 CEFR levels

	verb+direct combinations with object dependency	Paquot 2018	Types only: Non-significant increase from B2-C2 CEFR levels.
% medium-MI combinations	all bigrams	Granger & Bestgen 2014	Non-significant increase at than intermediate proficiency level for types.
(MI ≥5 and <7)			No difference in use between advanced and intermediate proficiency levels for tokens.
	premodifier+noun combinations	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types and tokens
	adjective+noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens.
	adjective+noun combinations with modifier dependency	Paquot 2018	Types only: Significant increase from B2-C2 CEFR levels.
	noun+noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens.
	adverb+adjective combinations	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
	adverb+adjective & adverb-verb combinations with modifier dependency	Paquot 2018	Types only: Non-significant increase between B2-C2 CEFR levels.
	verb+direct combinations with object dependency	Paquot 2018	Types only: Significant increase between B2-C2 CEFR levels.

	% low-MI combinations (MI ≥3 and <5)	all bigrams	Granger & Bestgen 2014	Used more at advanced than intermediate proficiency level but not statistically significant for types or tokens.
		premodifier+noun combinations	Granger & Bestgen 2014	Significantly more used at intermediate than advanced proficiency level for types (cohen's d=.60) and tokens (cohen's d=.61)
		adjective+noun bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced proficiency level for types (cohen's d=.65) and tokens (cohen's d=.66)
		adjective+noun combinations with modifier dependency	Paquot 2018	Types only: Significant linear increase between B2-C2 CEFR levels.
		noun+noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
		adverb+adjective combinations	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
		adverb+adjective & adverb-verb combinations with modifier dependency	Paquot 2018	Types only: Non-significant increase between B2-C2 CEFR levels.
		verb+direct combinations with object dependency	Paquot 2018	Types only: Non-significant increase from B2-C2 CEFR levels.
	% non-MI-collocation (MI < 3)	all bigrams	Granger & Bestgen 2014	Significantly more non-collocations used at intermediate than advanced proficiency levels for types (cohen's d = 1.01) and tokens (cohen's d = .88).

		Premodifier + noun bigrams	Granger & Bestgen 2014	Significantly more types used at intermediate than advanced proficiency level (cohen's d =.55) Significant more tokens used at intermediate level than advanced proficiency level (cohen's d=.42).
		adjective+noun bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced proficiency level for types (cohen's d-.51) and tokens (cohen's d=.38).
		Adjective + noun combinations with modifier dependency	Paquot 2018	Types only: Significant decrease between B2-C2 CEFR levels.
		Noun + noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
		Adverb + adjective bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced proficiency level for types (cohen's d=.29) and tokens (cohen's d=.27)
		adverb+adjective & adverb-verb combinations with modifier dependency	Paquot 2018	Types only: Non-significant decrease between B2-C2 CEFR levels
		verb+direct combinations with object dependency	Paquot 2018	Types only: Non-significant decrease between B2-C2 CEFR levels
Mean MI	Mean MI for all bigrams	MI based on occurrences in COCA	Bestgen & Granger 2014	Significant positive correlation with overall quality (types: r =.35; tokens: r=.28), language score (types: r=.43; tokens: r=.32) and vocabulary score (types: r=.31; tokens: r=.22)

			Bestgen 2017	Significant positive correlation with quality for FCE (r =.46) and ICLE (r =.60) texts.
		Bigram MI component made up of pooled MI for different COCA sub-corpora including spoken sub-corpora	Kim, Crossley & Kyle 2018	Significant positive correlation with writing proficiency scores (r=.20) and lexical proficiency scores (r=.21)
		MI based on occurrences in COCA academic (TAALES index)	Garner, Crossley & Kyle 2018	Significant increase between CEFR A2 and B2 proficiency levels (r=.24).
		MI based on occurrences in COCA spoken (TAALES index)	Garner, Crossley & Kyle 2018	Significant increase between CEFR A2 and B2 proficiency levels (r=.32).
	Mean MI for all trigrams	Trigram MI	Kim, Crossley & Kyle 2018	Significant negative correlation with writing proficiency scores (r =-.16). Significant positive correlation with lexical proficiency scores (r=.32)
		Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.13)
		MI based on occurrences in COCA academic Trigram (refers to association between: *word1* & *word2+word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.30)
		MI based on occurrences in COCA spoken Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.31)
		MI based on occurrences in COCA spoken	Garner, Crossley & Kyle 2019	Significant positive correlation with writing proficiency CEFR levels (r=.21).
Mean MI^2		MI^2 based on occurrences in COCA academic (TAALES index)	Garner, Crossley & Kyle 2018	Significant increase between CEFR A2 and B2 proficiency levels (r=.35).

	Mean MI^2 for all bigrams		Garner, Crossley & Kyle 2019	Significant positive correlation with writing proficiency CEFR levels (r=.27).
		MI^2 based on occurrences in COCA spoken (TAALES index)	Garner, Crossley & Kyle 2018	Significant increase between CEFR A2 and B2 proficiency levels (r=.22).
	Mean MI^2 for all trigrams	Trigram (refers to association between: *word1 & word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.16)
		MI^2 based on occurrences in COCA academic Trigram (refers to association between: *word1 & word2+word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.13).
		MI^2 based on occurrences in COCA spoken Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.16)
Mean MI per POS	Mean MI adjective + noun (identified by Stanford dependency parser)	MI^2 based on occurrences in COCA spoken Mean MI adjective + noun types	Paquot 2019	Statistically significant increases between CEFR groups (η^2 =.11) however Tukey post-hoc tests reveal Mean MI increases are only significant with B2-C2. B2- C1 and C1-C2 increases are non-significant.
			Paquot 2018	Statistically significant increase from B2-C2 CEFR levels.

	Mean MI adverbial modifier (adverb +adjective, adverb + verb) (identified by Stanford dependency parser)	Mean MI adverbial modifier types	Paquot 2019	Significant increase between adverb modifier measures across CEFR levels (η^2=.12) but Tukey tests reveal that this increase is only significant at B2-C1 and B2-C2 levels and not between C1-C2 CEFR levels.
			Paquot 2018	Statistically significant linear increase from B2-C2 CEFR levels.
	Mean MI verb + direct object (identified by Stanford dependency parser)	Mean MI verb + direct object types	Paquot 2019	Statistically significant increase between CEFR levels (η^2=.15) however Tukey post-hoc tests reveal that there are non-significant increases between B2-C1 CEFR levels and a statistically significant increase at B2-C2 and C1-C2 CEFR levels.
			Paquot 2018	Statistically significant increase from B2-C2 CEFR levels.
t-score threshold measures	% high t-score combinations (t ≥10)	All bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced level of proficiency for types (cohen's d=.33) and tokens (cohen's d=.43)
		Premodifier + noun bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced level of proficiency for types (cohen's d=.38) and (cohen's d=.43) for tokens
		noun + noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens

	adjective + noun bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced level of proficiency for types (cohen's d=.44) and tokens (cohen's d=.50)
	Adverb + adjective bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced level of proficiency for types (cohen's d=.45) and tokens (cohen's d=.47)
% medium t-score combinations (t ≥6 and < 10)	All bigrams	Granger & Bestgen 2014	Significantly more used at advanced than intermediate proficiency level for types (cohen's d=.64) and (cohen's d=.67) for tokens.
	Premodifier + noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
	noun + noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens
	adjective + noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
	Adverb + adjective bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
	All bigrams	Granger & Bestgen 2014	Significantly more used at advanced than intermediate proficiency level for

% low t-score combinations			types (cohen's d=.58) and (cohen's d=.56) for tokens
(t 2 and < 6)	Premodifier + noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens.
	noun + noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens
	adjective + noun bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens
	Adverb + adjective bigrams	Granger & Bestgen 2014	More used at advanced than intermediate proficiency level but not statistically significant for types or tokens
% t-score non-collocations (t < 2)	All bigrams	Granger & Bestgen 2014	Significantly more used at intermediate than advanced level of proficiency for types (cohen's d=.88) and (cohen's d=.71) for tokens
	Premodifier + noun bigrams	Granger & Bestgen 2014	Significantly more types used at intermediate than advanced proficiency level (cohen’s d=. 49) but not statistically significant for tokens
	noun + noun bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens

		adjective + noun bigrams	Granger & Bestgen 2014	Significantly more types used at intermediate than advanced proficiency level (cohen's d=.48) but not statistically significant for tokens
		Adverb + adjective bigrams	Granger & Bestgen 2014	More used at intermediate than advanced proficiency level but not statistically significant for types or tokens
Mean t-score	Mean t-score for all bigrams	t-score based on occurrences in COCA	Bestgen & Granger 2014	Weak non-significant correlation with overall quality, language and vocabulary scores for tokens and types.
		t-score based on occurrences in the BNC	Bestgen 2017	Significant positive correlation with FCE text quality (r=.10) but non-significant positive correlation with ICLE text quality (r=.03).
		t-score based on occurrences in COCA academic	Garner, Crossley & Kyle 2018	Significant increase across CEFR A2-B2 proficiency levels (r=.35).
	Mean t-score for all trigrams	Trigram (refers to association between: *word1 & word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.20).
		t-score based on occurrences in COCA academic Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.20).
Mean Delta P	Mean Delta P for all bigrams	t-score based on occurrences in COCA academic Delta P based on occurrences in COCA academic	Garner, Crossley & Kyle 2018	Significant increase across CEFR A2-B2 proficiency levels (r=.47)
			Garner, Crossley & Kyle 2019	Significant positive correlation with CEFR writing proficiency levels (r=.33).

		Delta P based on occurrences in COCA spoken	Garner, Crossley & Kyle 2018	Significant increase across CEFR A2-B2 proficiency levels (r=.35)
	Mean Delta P for all trigrams	Trigram (refers to association between: *word1 & word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.34).
		Delta P based on occurrences in COCA academic	Garner, Crossley & Kyle 2019	Significant positive correlation with writing proficiency CEFR levels (r=.23).
		Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.30).
		Delta P based on occurrences in COCA academic		
		Trigram (refers to association between: *word1 & word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.33).
		Delta P based on occurrences in COCA spoken		
		Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.36).
		Delta P based on occurrences in COCA spoken		
Mean collexeme score	Mean collexeme score for all bigrams	Collexeme score based on occurrences in COCA academic	Garner, Crossley & Kyle 2018	Significant increase across CEFR A2-B2 proficiency levels (r=.34)
	Mean collexeme score for all trigrams	Trigram (refers to association between: *word1 & word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.25).
		Collexeme score based on occurrences in COCA academic		

		Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=.18).
		Collexeme score based on occurrences in COCA academic		
		Trigram (refers to association between: *word1* & *word2+word3))*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=-.12).
		Collexeme score based on occurrences in COCA spoken		
		Trigram 2 (refers to association between: *word1+word2 &: word3)*	Garner, Crossley & Kyle 2018	Significant increase across A2-B2 CEFR proficiency levels (r=-.21).
		Collexeme score based on occurrences in COCA spoken		
Pooled bigram and trigram association measures	Bigram and Trigram strength of directional association (Delta P)	Delta P calculated by using different sub-corpora of COCA as reference corpora.	Kim, Crossley & Kyle 2018	Significant positive correlation with writing proficiency scores (r=.30) and lexical proficiency scores (r=.38).

Appendix D

IRB Letter of Approval

Date: 11 October 2018

To: Lee McCallum

From: Lisa Meloncon, MyR 2.0 System Coordinator, Assoc. Prof. Technical Communication

Re: **Responsibility and Acknowledgement of Information from *MyR 2.0*: Agreement**

I am writing to confirm that your research thesis with the working title: '*Relationships between collocations, assessment and rubric types and language status in determining writing quality in a US first year composition programme*' has been added to the standing USF IRB No. Pro00021265. We are pleased to add your research to the growing number of projects associated with our construct specific digital ecology for teaching and assessing writing.

So that we are able to continue offering information related to *MyR 2.0* to US and international researchers, we ask that you agree to the following by signing this letter and returning it to me.

- Student confidentiality: Your research has been approved for a specific design regarding textual analysis, and you have agreed to confidentiality of information within the database. Nevertheless, the text you will analyze may contain information that might be of a personal nature to the student writers. As such, you agree not to quote directly any detailed blocks of text that contain such information. If a block of text is required to support the claims made in your findings, you agree to submit that text to me for my review before peer review and subsequent publication.

- Data use and re-use: Your research has been approved for a specific research project. Should you want to use the data again in another capacity, please notify us. Relatedly, this data has been released to you and/or your research team. It is not to be distributed to any other parties for any reason.

- Acknowledgement: You agree to add the following statement to all presentations, grant proposals, conference proceedings, and publications based on information gained from our database: "Our data was compiled from the *MyR* databases hosted in the Department of English at the University of South Florida. Our research was approved under USF IRB No. Pro00021265."

In addition, you agree to provide copies of presentations, grant proposals, conference proceedings, and publications associated with the *MyR* database.

Please let me know if I can provide further support of your research.

Signature and date of researcher: Lee McCallum (11.10.2018)

Signature and date of My R and IRB coordinator Lisa Melancon 24 Oct 2018

Appendix E

Demographic Survey Questions

1. **What is your anticipated college major?**

   (Students choose from extensive drop down menu)

2. **How old are you?**

☐ Do not wish to answer
☐ Under 18 years old
☐ 21-24 years old
☐ 25-39 years old
☐ 40 years old or more

3. **Year in School**

☐ Do not wish to answer
☐ Freshman
☐ Sophomore
☐ Junior
☐ Senior
☐ Masters
☐ Doctoral
☐ Postdoctoral

4. **What is the first language that you learned?**

☐ Do not wish to answer
☐ English
☐ English + another language
☐ Mandarin
☐ Spanish
☐ Hindu/Urdu
☐ Arabic
☐ Portuguese
☐ Bengali
☐ Russian
☐ Japanese
☐ Punjabi
☐ Another language

5. **Gender**

☐ Male

☐ Female
☐ Transgender
☐ Other

6. Are you:

☐ Do not wish to answer
☐ Hispanic or Latino
☐ Not Hispanic or Latino

7. Indicate one or more races that apply among the following:

☐ Do not wish to answer
☐ American Indian or Alaska Native
☐ Asian
☐ Black or African American
☐ Native Hawaiian or Other Pacific Islander
☐ White

8a. Choose the highest level of education of one parent or guardian

☐ Do not wish to answer
☐ Don't know or unsure
☐ No high school diploma
☐ Associate degree
☐ Bachelors degree
☐ Graduate degree
☐ Professional degree (e.g. law or medical)

8b. Choose the highest level of education of a second parent or guardian

☐ Do not wish to answer
☐ Don't know or unsure
☐ No high school diploma
☐ Associate degree
☐ Bachelors degree
☐ Graduate degree
☐ Professional degree (e.g. law or medical)

9. What is the range of your family income?

☐ Do not wish to answer
☐ $0-20,000
☐ $20,001-40,000
☐ 40,001-60,000
☐ $60,001-80,000
☐ $80,001-100,000
☐ $100,001-120,000

☐ $120,001-140,000
☐ $140,001-160,000
☐ $160,001-200,000
☐ >$200,000

Appendix F

Annotator Agreement

The Role of Collocations and Learner and Course Variables in Determining Writing Quality in Assignments from a First-Year Composition Programme

My doctoral thesis examines the relationship between restricted collocations and writing quality as attested by essay grade scores from a corpus of coursework assignments obtained from a first-year composition programme in the U.S.

Collocation extraction uses the Stanford parser whereby the parser identifies collocations which have a dependency relationship. These collocations take the following syntactic relationships: adjective modified by noun, adverb modified by adjective, adverb modified with verb, nouns with subject dependency on a verb, nouns with an object dependency on a verb.

In attempt to maximise the validity and reliability of the research, my thesis will also evaluate the accuracy of the Stanford parser in correctly identifying these dependency relationships. In ensuring my own checking of the Stanford parser dependency relationships is accurate, I need an annotator to check the dependency relationships from the parser are in fact dependency relationships.

This work will require the annotator to check the Stanford parser output against the original texts and my own observations from the output. The work will require the annotator to check a range of texts that differ by their awarded grade. Based on the sample size (N=897), 10 texts at each grade level will be manually checked (n=180) and I would like to get 10% of this sample checked. This means the annotator would be responsible for checking 18 texts in total.

Based on my own experience, I would anticipate that each text would take around 30 minutes to check. Texts are around 1,000 words in length and so this judgement is based on the number of collocations that are realistically likely to be present in each text.

The second annotator is welcome to be included as a second author on publications that arise from the thesis.

Work is likely to start at the end of April 2019 and I request that the checks are completed by the end of May 2019. However, this is dependent on the schedule of the annotator.

Please note that the interested annotator should be familiar with learner texts presented in text files, MS Excel and be willing to sign the attached data protection agreement.

Thank you for your time.

Best wishes,

Lee McCallum

ANNOTATOR CONFIDENTIALITY AGREEMENT

This confidentiality agreement serves as an agreement between the project researcher (Lee McCallum) and the annotator who is appointed to the project.

By acting as an annotator, the annotator agrees to protect the data that is shared with them by acknowledging the following:

- Texts are to be accessed through a secure folder and **not** saved to their personal computer.
- Texts are **not** to be shared with any other party.
- Texts are **not** to be amended in any way.
- One copy of this form will be kept by the annotator; a second copy will be kept by the project researcher.

.....……..
.....
(Signature of annotator) (Date)

……
……
(Printed name of annotator) (Date)

.....……..
.....
(Signature of project researcher) (Date)

……
……
(Printed name of project researcher) (Date)

References

Ackermann, K., & Chen, Y.H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235-247.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *Journal of English for Specific Purposes*, *31*, 81-92.

Allen, D. (2010). Lexical bundles in learner writing: an analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education*, 1. 105–127.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie, T. (Ed.), *Comprehension and teaching: research reviews* (pp. 77-125). Newark, Delaware: International Reading Association.

Anthony, L. (2018). Suite of lexical tools [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software. Last accessed: 20.10.2018

Anson, C. M. (2000). Response and the social construction of error. *Assessing Writing*,*7*(1), 5–21. https://doi.org/10.1016/S1075-2935(00)00015-5

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high and low proficiency levels. *Language Assessment Quarterly*, *13*(1), 55-71.

Arthur, B. (1979). Short-term changes in EFL composition skills. In C. Yorio, K. Perkins, & J.Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 330-342). Washington, DC: TESOL.

Attali, Y., & Burnstein, J. (2006). Automated essay scoring with e-rater V2. *Journal of Technology, Learning and Assessment*, 4(3), 1-31.

Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, 1-29.

Aull, L. L. (2015a). *First-year university writing: A corpus-based study with implications for pedagogy*. UK: Palgrave Macmillan.

Aull, L.L. (2015b). Linguistic attention in rhetorical genre studies and first year writing. *Composition Forum*, 31.

Aull, L.L. (2015c). Connecting writing and language in assessment: Examining style, tone, and argument in the U.S Common Core standards and in exemplary student writing. *Assessing Writing*, *24*, 59-73.

Aull, L.L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. *Journal of Writing Analytics*, *1*, 1-47.

Aull, L.L. (2019). Linguistic markers of stance and genre in upper-level student writing. *Written Communication*, *36*(2), 267-295.

Baayen, H. (2008). Analysing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Barkaoui, K. (2008). Effects of scoring method and rater experience on ESL essay rating processes and outcomes. Unpublished PhD thesis. University of Toronto.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing*, *27*(4), 515-535.
Barlow, M. (2018) Collocate (version 2). Available from: http://www.michaelbarlow.com. Last accessed 20.10.2018.
Bartholomae, D. (1986). Inventing the university. In Rose, M. (Ed,). *When a writer can't write: Studies in writer's block and other composing problems* (pp. 134-165). Guilford Press, New York.
Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tubingen: Gunter Narr Verlag Tubingen.
Becker, A. (2010). Distinguishing linguistic and discourse features in ESL students' written performance. *Modern Journal of Applied Linguistics* 2, 406-424.
Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, *3*(1), 23–35.
Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes, 43*(2), 79-120.
Berman, R. A., & Nir, B. (2010). The lexicon in writing-speech-differentiation. *Written Language and Literacy, 13*(2), 183-205.
Berry-Rogghe, G (1973). The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (Eds.), *The Computer and Literary Studies* (pp. 103–112). Edinburgh.
Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.
Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65-78.
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405.
Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5-35.
Biber, D., & Gray, B.(2013). Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-grammatical Analysis. *TOEFL iBT Research Report* (TOEFL iBT-19). Educational Testing Service.
Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639-668.
Bouwer, R., Béguin, A., Sanders, T.J.M., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100.
Branham, C., Moxley, J., & Ross, V. (2015). My reviewers: participatory design and crowd-sourced usability processes. Proceedings of the 33rd Annual International Conference on the Design of Communication. July 16-17th Limerick, Ireland.
Brezina, V. (2018). *Statistics for corpus linguistics*. Cambridge: Cambridge University Press.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). V5.4 [software]. Retrieved, 16.03.2021, from: http://corpora.lancs.ac.uk/lancsbox

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1-22.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25,* 587-603.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

Bychkovska, T., & Lee, J.J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52.

Carlson, S., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Weiner, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 126-152). New York: Longman.

Carroll, J.B. (1964). *Language and thought*. Englewood cliffs, NJ: Prentice-Hall.

Carter, R., & McCarthy, M. (1988). *Vocabulary and language teaching*. London: Longman.

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing*, *3*, 179-201.

Celaya, M.L., & Naves, T. (2009). Age-related differences and associated factors in foreign language writing: Implications for L2 writing theory and school curricula. In R.M. Manchon (Ed.). *Writing in foreign language contexts: Learning, teaching and research* (pp. 2521-2961). Multilingual Matters.

Charney, D. (1984). The validity of holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* 65-81.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*, 30–49.

Chen, Y-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849-880.

Chen, W. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC Journal*, 1-18. https://doi.org/10.1177/0033688217716507

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling, (pp.609–623). Cambridge, USA.

Christensen, R.H.B., & Brockhoff, P.B. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de la Societe Francaise de Statistique & Revue de Statistique Appliquee*, *154*(3), 58-79.

Christensen R.H.B (2018). "ordinal—Regression Models for Ordinal Data." R package version 2018.8-25, Retrieved 26.06.2020, from, URL http://www.cran.r-project.org/package=ordinal/.

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, *16*, 22–29.

Church, K., Gale, W. A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Using on-line resources to build a lexicon* (pp. 115-164). Hillsdale, NJ: Lawrence

Erlbaum.
Clear, J. (1993). Tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp.271-292). Amsterdam: Benjamins Publishing Company.
Cobb, T. (2017/2019). Web Vocab profile. Available at: http://www.lextutor.ca/vpl, an adaptation of Heatley, Nation and Coxhead's (2002) Range. Last accessed: 24.02.2019.
Conference on College Composition and Communication (CCCC). (2014). Statement on second language writing and writers. Retrieved from http://www.ncte.org/cccc/resources/positions/secondlangwriting. Last accessed:24.02.2019.
Connors, R.J. (1997). *Composition-rhetoric: Backgrounds Theory and Pedagogy.* Pittsburgh: University of Pittsburgh Press.
Council of Writing Program Administration (CWPA). (2014). Outcomes statement for first-year composition (3.0). Retrieved from http://wpacouncil.org/positions/outcomes.html. Last accessed: 24.02.2019.
Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics, 2*, 223–235.
Cowie, A.P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M.McCarthy (Eds,). *Vocabulary and language teaching* (pp. 126-139). London: Longman.
Cowie, A.P. (1991). Multiword units in newspaper language. In S. Granger (Ed,). *Perspectives on the English lexicon: A tribute to Jacques van Roey* (pp. 101-116). Louvain -la-Neuve: Cahiers de l'institut de linguistique de Louvain.
Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), The encyclopedia of language and linguistics (pp. 3168–3171). Oxford, UK: Oxford University Press.
Cowie, A. P. (1998). Phraseology: Theory, analysis, and applications. Oxford: Oxford University Press.
Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly, 34*(2), 213-238.
Crawley, M.J. (2013). *The R book* (2nd edition). Wiley.
Credit – by - Exams Equivalent. (2018). Credit by Exams Equivalent: Undergraduate studies. Available at: http://ugs.usf.edu/credit-by-exam/ Last accessed: 24.02.2019
Creswell, J. (2014). *Research design: Qualitative, quantitative and mixed methods approaches*. (4th edition). London: Sage.
Crossley, S.A., & McNamara, D.S. (2009). Computational assessment of lexical differences in L1 and L2 writing, *Journal of Second Language Writing*, *18*,119-135.
Crossley, S.A., & McNamara, D.S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In R. Catrambone & S, Ohlsson (Eds.). Proceedings of the 32nd conference of the Cognitive Science Society (pp. 984-989). Austin, TX: Cognitive Science Society. Available at: http://csjarchive.cogsci.rpi.edu/Proceedings/2010/papers/0310/paper0310.pdf. Last accessed: 8/9/2016.
Crossley, S.A., & McNamara, D.S. (2011). Understanding expert ratings of

essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning 21*(2), 170-191.
Crossley, S.A., & McNamara, D.S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115-135.
Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580.
Crossley, S.A., Weston, J.L., Sullivan, S.T.M., & McNamara, D.S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282-311.
Crossley, S.A., Cai, Z., & McNamara, D.S. (2012). Syntagmatic, paradigmatic and automatic n-gram approaches to assessing essay quality. Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference (pp. 214-219). Palo-Alto, CA: The AAAI Press.
Crossley, S. A., Defore, C., Kyle, K, Dai, J., & McNamara, D. S. (2013). Paragraph specific N-Gram approaches to automatically assessing essay quality. In D'Mello, S. K., Calvo, R. A., & Olney, A. (Eds.) *Proceedings of the 6th Educational Data Mining (EDM) Conference.* (pp. 216-220). Heidelberg, Berlin, Germany: Springer.
Crossley, S.A., Salsbury, T., & McNamara, D.S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36* (5), 570-590.
Crossley, S.A., Kyle, K., & McNamara, D.S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essays quality. *Journal of Second Language Writing*, *32*, 1-16.
Crotty, M. (2015). *The foundations of social research*. London: Sage.
Cruse, D. A. (1986). *Lexical semantics*. New York: Cambridge University Press.
Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5-43.
Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontenous speech of bilinguals. *Applied Linguistics*, *24*(2), 197-222.
Daller, H., Turlik, J., & Weir, I. (2013). Vocabulary acquisition and the learning curve. In S. Jarvis & H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins.
Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*(2), 151-177.

Deville, G. & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A.Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics:Multiple perspectives* (pp. 9-25). Amsterdam: John Benjamins.

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*(4), 241-256.

Downey, B., Loyer, E., Walkup, K. (2016). (Eds,). *The MyReviewers guide to style and grammar*. Florida: University of South Florida.

Douglas, S.R. (2015). The relationship between lexical frequency profiling measures and rater judgements of spoken and written general English language proficiency on the CELPIP-General test, TESL Canada, *32*(9), 43-64.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics in Language Teaching, 47, 157–177.

Durrant, P. (2017).Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*, 165-193.

Durrant, P. (2019). Formulaic language in English for Academic Purposes. In *Understanding formulaic language: A second language acquisition perspective* (pp. 211-228). London: Routledge.

Durrant, P. (2020). Association measure calculator. Retrieved from: https://phildurrant.net/resources/ Last accessed: 27.07.2020

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in freshman composition assignments. *International Journal of Corpus Linguistics*, *24*(1), 31-64.

Durrant, P. & Brenchley, M. (in press). The development of academic collocations in children's writing. To appear In P. Szudarski & S. Barclay (Eds,). *Vocabulary Theory, Patterning and Teaching*. Bristol: Multilingual Matters.

Durrant, P., Brenchley, M., & McCallum, L. (2021). *Understanding development and proficiency in writing: Quantitative corpus linguistics approaches*. Cambridge: Cambridge University Press.

Eager, C., Roy, J. (2017). Mixed Effects Models are Sometimes Terrible. Retrieved from. http://arxiv.org/abs/1701.04858.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 texts and writers in first-year composition, *TESOL Quarterly*, 52(1),137-162.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, N.C. (2008). Phraseology: The periphery and the heart of language. In: F. Meunier, & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 1–13). Amsterdam / Philadelphia, PA: Benjamins.

Ellis, N.C., & Vlach-Simpson, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61-78.

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139-155.
Espada-Gustilo, L. (2011). Linguistic features that impact essay scores: A corpus linguistic analysis of ESL writing in 3 proficiency levels. *The Southeast Asian Journal of English Language Studies*, *17*(1), 55-64.
Everitt, B.S. (1993). *Cluster analysis*. New York: John Wiley & Sons.
Evert, S. (2004). The statistics of word cooccurrences: Word pairs and collocations. (Doctoral dissertation) Stuttgart, Germany: University of Stuttgart.
Evert, S. (2007). Corpora and collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin.
Evert, S., and Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.
Ferris, D.R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*(2), 414-420.
Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
Finch, W.H., Bolin, J.E., & Kelley, K .(2014). *Multilevel modelling using R*. Florida, CRC Press.
Firth, J. R. (1957). *Papers in linguistics 1934–1951*. Oxford: Oxford University Press.
Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected papers of J.R. Firth 1952-1959* (pp. 168-205). Harlow: Longman.
Fletcher, W.H. (2002-2005). *KfNgram*. Annapolis, MD: USNA.
Framework for Success in Postsecondary Writing. (2011). Available from: http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf. Last accessed: 10.07.2019.
Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing and interpreting the evidence. *Language Learning*, *67*, 155-179.
Gardner, S., Nesi, H., & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, *40*(4), 646-674.
Garner, J.R. (2016). A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research*, *2*(1), 31-68.
Garner, J., Crossley, S., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of ngrams: An association measures study. *International Review of Applied Linguistics*, Ahead of print. DOI: https://doi.org/10.1515/iral-2017-0089
Garner, J., Crossley, S., & Kyle, K. (2019). Ngrams and L2 writing proficiency. *System*, 1-37. DOI: 10.1016/j.system.2018.12.001
Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Databases (EFCAMDAT). In Proceedings of the 31st Second

Language Research Forum: Building bridges between disciplines. Somerville, Cascadilla. Proceedings Project.
Gere, A. R. (1980). Written composition: Toward a theory of evaluation. *College English, 42,* 44-48.
Gere, A.R. (2016). A new perspective on language – level writing instruction. *Writing Program Administration*, *39*(2), 140.145.
Gilquin, G. (2017). *POS tagging a spoken learner corpus_Testing accuracy testing*. Paper presented at the Learner Corpus Research Conference, Bozen/Bolzano, Italy.
Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.
Granger. G., & Paquot M. (2008). Disentangling the phraseological web. In: Granger. S and Meunier F. (Eds,). *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam / Philadelphia, PA: John Benjamins Publishing Company.
Granger, G., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles., D. Pecorari., & S. Hunston (Eds,). *Academic writing: At the interface of corpus and discourse.* London: Continuum.
Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics*, 52(3), 229-252.
Grant, L & Ginther, A. (2000). Using computer tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123-145.
Greener, I. (2011). *Designing social research: A guide for the bewildered*. London: Sage.
Gries, S.Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier. (Eds.). *Phraseology: An interdisciplinary perspective* (pp. 27-49). Amsterdam/Philadelphia, PA: Benjamins.
Gries, S. Th. (2013a). 50-something years of work on collocations: What is or should be next ….*International Journal of Corpus Linguistics, 18*(1). 137–165.
Gries, S.Th. (2013b). Statistics for linguists with R: A practical introduction (2nd revised edition). Berlin: de Gruyter Mouton.
Gries, S.Th. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed effects) models. *Corpora*, *10*(1), 95-126.
Gries, S.Th. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1*(2), 276-307.
Gries, S.Th., & Durrant, P. (2020). Analysing co-occurrence data. In S. Gries & M. Paquot (Eds). *A practical handbook of corpus linguistics,* (pp. 141-159). New York: Springer.
Grix, J. (2004). *The foundations of research*. London: Palgrave MacMillan.
Guo, L., Crossley, S.A., McNamara, D.S. (2013). Predicting human judgements of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218-238.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners.* (PhD). Lund University, Lund.
Gyllstad, H., Granfeldt, J., Bernardini, P., & Kallkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, *14*, 1-30.
Gyllstad, H., & Wolter, B. (2015). Collocational processing in the light of a phraseological continuum model. Does semantic transparency matter? *Language Learning*,
Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies (pp. 153-167). New York: Longman
Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday., & R.H. Robins (Eds.), In memory of J. R. Firth (pp. 148–162). London: Longman.
Halliday, M.A.K. & Hasan, R. (2001). *Cohesion in English*. Beijing: Foreign Language Teaching and Research Press.
Hamp-Lyons, L. & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3,* 85-96.
Harrison, J., & Barker, F. (2015). (Eds.,). *English Profile Studies: English Profile in Practice*. Cambridge: Cambridge University Press.
Hausmann, F.J. (1985). Kollokationen im deutschen w¨orterbuch. ein beitrag zurtheorie des lexikographischen beispiels. In H. Bergenholtz & J. Mugdan, (Eds,)., Lezikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatikim W¨orterbuch, Lexicographica. Series Major *3*, (pp. 118–129) .
Hawkins, J.A., & Filipovic, L. (2012). *Criterial features in L2 English: Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.
Heck, R.H., & Thomas, S.L. (2000). *An introduction to multilevel modelling techniques*. New Jersey, NJ: Lawrence Erlbaum Associates.
Henriksen, B. (2013). Research on L2 learners' collocational competence and development: A progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis (pp. 29-56). Amsterdam: EUROSLA.
Hirose, K., & Sasaki, M. (1994). Explantory variables for Japanese students' expository writing in English: An exploratory study. *Journal of Second Language Writing*, *3*(3), 203-229.
Hillocks, G. (2002). The Testing trap: How states writing assessments control learning. New York: Teachers College Press.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
Hoffman, A., & Wiggs, K. (2016a). (Eds,). *Rhetoric matters: Foundations of rhetoric and composition*. Acton, MA: XanEdu. E-text.
Hoffman, A., & Wiggs, K. (2016b). (Eds,). *Rhetoric really matters: Explorations and argumentation*. Acton, MA: XanEdu.
Horst, M., & Collins, L. (2006). From fallible to strong: How does their vocabulary grow? *Canadian Modern Language Review*, *63*(1), 83-106.
Hosmer, D.W., & Lemeshow, S. (2010). *Applied logistic regression.* Wiley.

Hou, J., Loerts, H., & Verspoor, M. (2016). Chunk use and development in advanced Chinese L2 learners of English. *Language Teaching Research*, 1-21.
Hou, J., Verspoor, M., & Loerts, H. (2016). An exploratory study into the dynamics of Chinese L2 writing development. *Dutch Journal of Applied Linguistics*, *5*(1), 65-96.
Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in Second Language Acquisition.* Amsterdam: John Benjamins Publishing Company.
Howarth, P. (1996). Phraseology in English academic writing: Some implications for Language Learning and Dictionary Making. Tubingen: Niemeyer.
Howarth, P. (1998a). Phraseology and second language proficiency. *Applied Linguistics*, *19* (1), 24-44.
Howarth, P. (1998b). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), Phraseology: Theory, analysis, and applications (pp. 161–186). Oxford: Oxford University Press.
Howell, K.E. (2013). *The Philosophy of Methodology*. London: Sage.
Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
Huang, J., & Foote, C.J. (2010). Grading between lines: what really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, *7*(3), 219- 233.
Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*(1), 28-54.
Huot, B. A. (1990a). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* 201-213.
Huot, B. (1990b). The literature on direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237-263.
Hunt, K.W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly*, *4*(3), 195-202.
Hyland, K., & Guinda-Sancho, C. (2012) (Eds,). *Stance and voice in written academic genres.* London: Palgrave Macmillan.
Hyland, K. (2013). Writing in the university: Education, knowledge and reputation. *Language Teaching*, *46*(1), 53-70.
Izenman, A.J. (2013). *Modern multivariate statistical techniques: Regression, classification and manifold learning*. New York: Springer.
Jarvis, S. (2002). Short-texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57-84.
Jeffery, J.V., & Wilcox, K.C. (2013). How do I do it if I don't like writing? : Adolescents' stance toward writing across disciplines. *Reading & Writing*, *27*(6), 1095-1117.
Jespersen, O. 1974/1976. Living grammar. In The philosophy of grammar, 17–29. London: George Allen and Unwin. Reprinted in Diane D. Bornstein (ed.), *Readings in the theory of grammar*. (pp. 82–93). Cambridge, MA: Winthrop Publishers.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guildford Press.
Jones, S., & Sinclair, J. McH. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de Lexicologie, 24*, 15-61.
Kameen, P.T. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins, & J. Schachter. (Eds.). *On TESOL '79: The learner in focus* (pp. 343-364). Washington, D.C: TESOL.
Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study, *English Teaching*, *69*(4), 27-51.
Kim, M., Crossley, S.A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development and writing quality. *The Modern Language Journal*, *102*(1), 120-141.
Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Clarendon Press.
Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, *21*, 1-17.
Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, *28*, 39-52.
Kobayashi, H., & Rinnert, C. (2013). L1/L2/L3 writing development: Longitudinal case study of a Japanese multicompetent writer. *Journal of Second Language Writing*, *22*, 4-33.
Krenn, B. (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations.*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI & Universität des Saarlandes, Saarbrücken, Germany.
Krzeminska-Adamek, M. (2016). Lexis in writing: Investigating the relationship between lexical richness and the quality of advanced learners' texts. In M. Pawlak (Ed.), *Classroom-oriented research: Reconciling theory and practice* (pp. 195-197). Switzerland: Springer.
Kyle, K. (2017). Modelling quality in source-based texts. Available at: https://a4li.sri.com/archive/papers/Kyle_2017_Writing_Quality.pdf Last accessed: 28/07/2017.
Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings and application. *TESOL Quarterly*, *49* (4), 757-786.
Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of second language writing, 34*, 12-24.
Kyle, K., & Crossley, S.A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, *102*(2), 333-349.
Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*, 21-33.
Laufer, B., & Nation, P. (1995). Vocabulary size and use lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.
Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255-

271.Lawton, D. (1963). Social class differences in language development: A study of some samples of written work. *Language and Speech, 6*(3), 120-143.
Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647-672.
Lavallee, M., & McDonough, K. (2015). Comparing the lexical features of EAP students' essays by prompt and rating. *TESL Canada*, *32* (2), 31-44.
Lee, H-J. (1992). Measures of quality in L1 and L2 writing by two ESL learners. *English Teaching*, *44*, 89-113.
Lee, J. (2019). A comparison of writing tasks in ESL writing and first-year composition courses: A case study of one university. *Language Teaching Research*, 1-18.
Leki, I., & Carson, J.G. (1994). Students' perceptions of EAP writng instruction and writing needs across the disciplines. *TESOL Quarterly*, *28*(1), 81-101.
Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies*, *7*(3), 163-180.
Levitzky-Aviad, T. (2012). Lexical richness and variation in the writing of school-age EFL learners at different learning stages and different educational systems". In Y. Tono, Y. Kawaguchi & M. Minegishi (Eds.), *Developmental and Cross-linguistic Perspectives in Learner Corpus Research* (pp. 159-168). Amsterdam: John Benjamins.
Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over 8 years of study: Single words and collocations. In C. Bardel., C. Lindquist., & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp.127-148). *EUROSLA Monograph Series 2.*
Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Lewis, M. (1993), *The Lexical Approach*, Hove: Language Teaching Publications.
Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory Into Practice*. Hove: Language Teaching Publications.
Lewis, M. (Ed,). (2000). *Teaching collocations: further developments in the lexical approach*. Boston: Thomson.
Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, 85-102.
Li, X. (1996). *Good writing in cross-cultural contexts*. Albany, NY: SUNY Press.
Liu, X. (2016). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling*. Thousand Oaks, CA: Sage.
Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Sweden: Gleerup.
Llach, M.P.A. (2007). Lexical errors as writing quality predictors. *Studia Linguistica, 6*(1), 1-19.
Lorenz, G. (1999). *Adjective intensification - learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as

indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208.
Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27.
Luke, A.D. (2011). Basic multilevel modelling. In *Multilevel modelling* (pp. 10-53). London: Sage
Malvern, D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development*. Basingstoke: Palgrave Macmillan.
Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.
Marco, M.J.L. (2011). Exploring atypical verb + noun combinations in learner technical writing. *International Journal of English Studies*, *11*(2), 77-95.
Matsuda, P.K. (2006). The myth of linguistic homogeneity in U.S college writing, *College English*, *68*(6), 637-651.
Matsuda, P. K. (2012). Let's face it: Language issues and the writing program administrator. *Writing Program Administration*, *36*(1), 141–163.
Matsuda, P. K., Saenkhum, T., & Accardi, S. (2013). Writing teachers' perceptions of the presence and needs of second language writers: An institutional case study. *Journal of Second Language Writing*, *22,* 68–86. https://doi.org/10.1016/j.jslw.2012.10.001
Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3-15.
McCallum, L. (2020). R scripts and output. Retrieved from, https://leemccallum.net/resources/ Last accessed 29.09.2020.
McCoach, B. (2010). Hierarchical linear modeling. In G.R. Hancock., & R.O. Mueller, (Eds,). *The reviewers guide to quantitative methods in social sciences* (pp.123-141). London: Routledge.
McIntosh, C., Francis, B., and Poole, R. (2009). *Oxford Collocations Dictionary for students of English*. Oxford: Oxford University Press.
Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), Performance and competence in second language acquisition (pp. 35–53). Cambridge, England: Cambridge University Press.
Meara, P., & Bell, H. (2001). P-Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5-19.
Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: opportunities for collaboration with computational linguistics. *Language Learning*, *67*:S1, 66-95.
Michigan Corpus of Student Upper Papers (MICUSP). Retrieved from https://micusp.elicorpora.info/
Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*, 325–342.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4(*1), 43– 66.
Moxley, J.M. (2012). Aggregated assessment and objectivity 2.0. Proceedings of the EACL 2012. Workshop on computational linguistics and writing, (pp. 19-26). Avignon, France, April 23.
Moxley, J. (2013). Big data, learning analytics and social assessment. *The Journal of Writing Assessment*, 6 (1), 1-12.
Moxley, J.M., & Eubanks, D. (2015). On keeping score: Instructors vs. students; rubric ratings of 46, 689 essays. *Writing Program Administration*, *39*(2), 53-80.
My Reviewers Student Manual. (2018). Available at: https://enc1101fall12.files.wordpress.com/2012/09/ver41_student_manual_myreviewers.pdf Last accessed: 24.02.2019.
Myles, F. (2004). From data to theory: The over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society*, *102*, 139–168.
Myles, F. (2012). Complexity, accuracy and fluency: The role played by formulaic sequences in early interlanguage development. In A. Housen, F. Kuiken., & I. Vedder (Eds.). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in Second Language Acquisition* (pp.71-93). Amsterdam: John Benjamins.
Myles, F., & Cordier, C. (2017). Formulaic sequences (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition*, *39*, 3-28.
Nation, P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nattinger, J. R., & DeCarrico, J. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.
Negretti, R. (2012). Metacognition in Student Academic Writing. *Written Communication, 29*(2), 142– 179. doi: 10.1177/0741088312438529.
Nesselhauf, N. (2005). Collocations in a learner corpus. *Studies in Corpus Linguistics*, *14*. Amsterdam: John Benjamins Publishing Company.
Nihalani, N.K. (1981). The quest for the L2 index of development. *RELC Journal*, *12*(2), 50-56.
Nosratinia, M., & Razavi, F. (2016). Writing complexity, accuracy and fluency among EFL learners: Inspecting their interaction with learners' degree of creativity. *Theory and Practice in Language Studies*, *6*(5), 1043-1052.
North, B. (2014). *English Profile Studies: The CEFR in Practice*. Cambridge: Cambridge University Press.
O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.
O'Connell, A.A. (2010). An illustration of multilevel models for ordinal response data. In C. Reading (Ed.), Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php
O'Donnell, M.B., Römer, U., & Ellis, N.C. (2013). The development of formulaic sequences in first and second language writing, *International Journal of Corpus Linguistics*, *18*(1), 83-108.

Olinghouse, N., G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*, 45-65.
Onwuegbuzie, A.J., Johnson, R.B., Collins, K.M.T. (2009). Call for mixed analysis: A philosophical framework for combining qualitative and quantitative approaches, *International Journal of Multiple Research* 3(2), 114-139.
Osbourne, J.W. (2010). Correlation and other measures of association. In G.R. Hancock., & R.O. Mueller, (Eds,). *The reviewers guide to quantitative methods in social sciences* (pp. 55-71). London: Routledge.
Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6,* 1-13.
Osgood, C.E. (1952). The nature and measurement of meaning. *Psych Bull*, *49*, 197 - 237.
Osgood, C.E. (1957). The measurement of meaning. University of Illnois Press, Urbana, Illnois.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601.
Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*, 117-134.
Palmer, H. (1933/1966), *Second Interim Report on English Collocations*, Tokyo: Kaitakusha. Reprinted 1966.
Pastor, D. (2010). Cluster analysis. In G.R. Hancock., & R.O. Mueller, (Eds,). *The reviewers guide to quantitative methods in social sciences* (pp. 41-55). London: Routledge.
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130-149.
Paquot, M. (2018). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121-145.
Paquot M. (2019). Phraseological competence: a useful toolbox to delimitate CEFR levels in higher education? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, 15(1), 29-43.
Pawley, A & Syder. F. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt (Eds,). *Language and Communication*, 191–226. London: Longman.
Pecina, P. (2005). An extensive empirical study of collocation extraction methods. Proceedings of the ACL Student Research Workshop, pp. 13-18. Ann Arbor, Michigan.
Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, *44*, 137-158.
Pecina, P., & Schlesinger, P. (2006). Combining association measures for collocation extraction. In Proceedings of the 21st International Conference Computational Linguistics and 44th annual meeting of the association for Computational Linguistics (COLING/ACL, 2006). Sydney, Australia.
Pedersen, Ted (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX.
Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education, 28, 56-78.* doi: 10.1007/s40593-016-0122-z

Points of Pride USF. (2018). Points of Pride USF. Available at: https://www.usf.edu/about-usf/points-of-pride.aspx. Last accessed: 24.02.2019

Polio, C., & Friedman, D.A. (2016). *Understanding, evaluating and conducting second language writing*. London: Routledge.

Pring, R. (2014). *Philosophy of educational research*. New York: Bloomsbury.

Punch, K.F., & Oancea, A.E. (2014). *Introduction to research methods in education* (2nd edition). London: Sage.

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners' writing performance. *Journal of Second Language Writing*, *33*, 3-17.

Quellmalz, E. S., Capell, F. J., & Chov, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19,* 241-258.

R Core Development Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria. URL http://www.R-project.org: R Foundation for Statistical Computing.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A.Kunnan (Ed.), Validation in language assessment (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Regulation 8.005. (2018). Academic policies and procedures. Available at: http://ugs.usf.edu/pdf/cat1819/8-academic-policies-and-procedures.pdf . Last accessed: 24.02.2019.

Reid, S., & Findlay, G. (1986). Writer's workbench analysis of holistically scored essays. *Computers and Composition*, *3*(2), 6-32.

Riazi, A.M. (2016). Comparing writing performance in TOEFL-Ibt and academic assignments: An exploration of textual features. *Assessing Writing*, *28*, 15-27.

Richards, J. C., & Renandya, W. A. (2002). *Methodology in language teaching: An anthology of current practice.* Cambridge: Cambridge University Press.

Ringler, H., Klebanov, B.B., & Kaufer, D. (2018). Placing writing tasks in local and global contexts: The case of argumentative writing. *The Journal of Writing Analytics*, *2*, 34-77.

Römer, U. (2009). English in academia: Does nativeness matter? Anglistik: *International Journal of English Studies*, *20*, 89–100.

Römer, U. (2010). Establishing the Phraseological Profile of a Text Type: The Construction of Meaning in Academic Book Reviews. *English Text Construction*, *3*(1), 95-119.

Römer, U., & O'Donnell, M.B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*(2), 159-177.

Rowbottom, D.P., & Aiston, S.J. (2006). The myth of Scientific Method in contemporary educational research. *Journal of Philosophy of Education*, *40*(2), 137-156.

Ruan, Z. (2016). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC Journal*, 1-14.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing, *TESOL Quarterly*, *45* (1), 63-80.
Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Ablex Publishing Corporation. Norwood, New Jersey.
Santorini, B. (1990). Part-of-Speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). Accessed at: https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf. Last accessed: 07.01.2020
Santos, T. (1988). Professors' reactions to the writing of non-native speaking students. *TESOL Quarterly*, *22*(1), 69-90.
Santos, D.O-Victor., Verspoor, M., & Nerbonne, J. (2013). Identifying important factors in essay grading using machine learning. In D. Tsagari., S. Papadima-Sophocleous., & S. Ioannou-Georgiou (Eds,). *International Experiences in Language Testing and Assessment. Selected Papers in Memory of Pavlos Pavlou* Vol 28, (pp. 295-309). Peter Lang: Frankfurt. *Language Testing and Evaluation*, Vol. 28.
Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, *46*(1), 137-174.
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.
Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language learning* (2nd edition). Cambridge: Cambridge University Press.
Schneider, U. (2018). ΔP as a measure of collocation strength: Considerations based on analyses of hesitation placement in spontaneous speech. *Corpus Linguistics and Linguistic Theory*,
Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22,* 1-30.
Scott, M. (2018). WordSmith Tools version 7, Stroud: Lexical Analysis Software. Available at: https://lexically.net/wordsmith/ Last accessed: 20.10.2018.
Seretan, V. (2011). *Syntax-based collocation extraction: Text, speech and language technology series* (Vol. 44). Geneva, Switzerland: Springer Science and Business Media. doi:10.1007/978-94-007-0134-2_4
Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: ESL research and its implications. *TESOL Quarterly*, *27*(4), 657- 677.
Sinclair, J. M. (1987). Collocation: a progress report. In R. Steele & T. Threadgold (Eds.), *Language topics: Essays in honour of Michael Halliday* (Vol. 2, pp. 319-331). Amsterdam: John Benjamins.
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), Contrastive lexical semantics (pp. 1–24). Amsterdam: John Benjamins Publishing Company.
Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics & Linguistic Theory*, *11*(2), 285-301.
Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2019). Formulaic language: setting the scene. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language* (pp. 1-15). London: Routledge.
Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), 143–177.

Song, C. B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5,* 163-182.

Spaan, M. (1993). The effect of prompt on essay examinations. In D. Douglas & C.Chapelle (Eds.), *A new decade of language testing research* (pp. 98-122).Alexandria, VA: TESOL.

Spurling, J. (2014). Evaluating lexical quality in writing in first and second language learners. Unpublished MA thesis, University of Victoria.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*, 214-225.

Staples, S., & Reppen, R. (2016). Understanding first – year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, *32*, 17-35.

Stevens, M.E., Giuliano, V.E. & Heilprin, L.B., (1964) (Eds,). *Statistical Association Methods for Mechanised Documentation*, National Bureau of Standards, Washington.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language*, *18*, 103-118.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207-223.

Sudweek, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261.

Tabachnick, B.G., & Fidell, L.S. (2014). *Using multivariate statistics* (6th edition).Harlow, UK: Pearson Education Limited.

Taguchi, N., Crawford, W., & Wetzel, D.Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, 47 (2), 420-430.

Tan, P-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining*. Harlow, Essex: Pearson Education Limited.

Tang, R. (2012). *Academic writing in a second or foreign language: Issues and challenges facing ESL/EFL academic writers in higher education contexts*. London: Continuum.

Tang, X., & Cao, J. (2015). Automatic genre classification via N-grams of Part-of-Speech Tags, 7th conference on Corpus Linguistics: Current work on corpus linguistics: Working with traditionally-conceived corpora and beyond (CILC 2015). *Procedia – Social and Behavioral Sciences*, *198*, 474-478.

Tedick, D. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*, 123-143.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Bejamins Publishing Company.

Treffers-Daller, J.T., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 39(3), 302-327.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education, 28*, 79-105.

Vann, R.J. (1979). Oral and written syntactic relationships in second language learning (pp. 322-330). In C. Yorio, K. Perkins, & J. Schachter. (Eds.). *On TESOL '79: The learner in focus* (pp.322-330). Washington, D.C: TESOL.

Van Rooy, B., & Schafer, L. (2003). Automatic POS tagging of a learner corpus: The influence of learner errors on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Vol. Eds.), Technical Papers 16: Proceedings of the Corpus Linguistics 2003 Conference (pp. 835–844). Lancaster, UK: Lancaster University Centre for Computer Corpus Research on Language.

Van Rooy., & Schafer, L. (2009). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, *20*(4), 325-335.

Verspoor, M., Schmid, S.M., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239-263.

Verspoor, M., Lowie, M., Ping - Chan, H., & Vahtrick, L. (2017). Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches en didactique des langues et des cultures*, *14*(1), 1-28.

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life writing examinations. *Research Notes*, *41*, 7-14.

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL*, *9*, 65-88.

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. Assessing Writing, 6, 145–178.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, *4*(2), 253-290.

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151-171.

Winter, B. (2020). *Statistics for linguists: An introduction using R*. New York: Routledge.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2006). Formulaic language. In. K. Brown (Ed,). *Encyclopedia of language and linguistics*, Volume 4 (pp. 590-597). Oxford: Elsevier.

Wray, A. (2008). Formulaic Language: Pushing the Boundaries. Oxford University Press.

Wray, A. (2019). Concluding question: Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language* (pp. 248-269). London: Routledge.

Wray, A., & Perkins, M.R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*, 1-28.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). Second language

development in writing: Measures of fluency, accuracy, and complexity. Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.
Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. London: Bloomsbury Publishing Plc.
Xiao, R. (2015). Collocation. In D. Biber., & R. Reppen (Eds,). *The handbook of corpus linguistics*. Cambridge: Cambridge University Press.
Yoon, H-J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing*, *34*, 42-57.
Yoon, H-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality*. System, 66*, 130-141.
Yoon, H.**,** & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51,* 275-301.
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31* (2), 236 - 259.
Zawacki, T.M., & Habib.A.S. (2014). Negotiating 'Errors' in L2 Writing: Faculty Dispositions and Language Difference. In T.M, Zawacki, & M. Cox. (Eds,). WAC and Second Language Writers: Research Towards Linguistically and Culturally Inclusive Programs and Practices (pp.183-210). Fort Collins, CO: The WAC Clearinghouse and Parlor Press.
Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English, *System*, *56*, 40-53.
Zhu, W. (2004). Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of Second Language Writing*, *13*, 29-48.
Zuur, A.F., Leno, E.N., Walker, N. & Saveliev, A.A. (2009). *Mixed effects models and extensions in ecology with R*. Berlin and New York: Springer.