2014

Understanding the Effect of Formulaic Language on ESL Teachers' Perceptions of Advanced L2 Writing: An Application of Corpus-Identified Formulaic Language

Alison Youngblood
University of Central Florida

UNDERSTANDING THE EFFECT OF FORMULAIC LANGUAGE ON ESL TEACHERS' PERCEPTIONS OF ADVANCED L2 WRITING: AN APPLICATION OF CORPUS-IDENTIFIED FORMULAIC LANGUAGE

by

ALISON M. YOUNGBLOOD

B.S. Florida State University, 2003
M.A. University of Central Florida, 2007

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Summer Term
2014

Major Advisors: Keith S. Folse and Joyce W. Nutta

ABSTRACT

A quantitative study was conducted to determine if the amount of formulaic language influenced ESL teachers' perceptions *(n=102)* of non-native writing skill, as evidenced by composite and sub-scale scores on the ESL Composition Profile (Jacobs et al., 1981). Formulaic language was operationalized as 25 three-word strings sampled from the writing sub-list of the Academic Formulas List (Simpson-Vlach & Ellis, 2010) and further validated as frequent in the Michigan Corpus of Upper Level Student Papers. The target formulaic sequences were divided into three experimental groups representing a low, mid, and high amount of formulaic language. Four advanced non-native writers generated argumentative, timed writing samples that incorporated the target sequences. The writing samples were then assembled into data collection packets and distributed at eight Intensive English Programs across the southeastern United States.

A repeated measures ANOVA indicated that there was a significant difference in composite score ($p < .05$) between the control and three experimental conditions; however, the essays that incorporated 16 and 25 formulaic sequences scored significantly lower than those with zero or eight target sequences. When the amount of syntactical and semantic errors were strictly controlled for, the composite scores also fell between the control and experimental conditions, but the decrease in score was not significant ($p > .05$).

The content, organization, vocabulary, language, and mechanics sub-scales were also compared using a repeated measures MANOVA. In content, organization, and language, the control and low essays outscored the mid and high conditions ($p < .05$). For the vocabulary sub-scale, the control and low condition were not significantly different, but the control essays

only outperformed the mid level essays. The low essays outperformed both the mid and high essays. In terms of mechanics, there was only a significant difference between the low and mid level essays. The results of the MANOVA were consistent when the amount of syntactic and semantic errors were controlled.

Implications for teaching suggest that the Academic Formulas List would not benefit academically-oriented L2 learners preparing to enter a university. While corpus tools are valuable in helping teachers, material writers, and publishers improve vocabulary instruction in the English classroom, not all statistically salient lexical combinations are important for non-native writers to master and incorporate in their academic discourse.

In memory of Mary Lord

ACKNOWLEDGMENTS

The completion of a doctoral degree is in no way a solo activity. The culmination of this endeavor was made possible through the support of family, friends, mentors, and colleagues, and as a thank you, I offer the following simple words of gratitude.

I would like to thank Dr. Keith Folse, Co-chair of my dissertation, for his support and encouragement. I am eternally grateful for the countless hours he has spent revising drafts, answering emails, hashing out details, and even traveling with me for data collection. However, Dr. Folse's hand in my successful completion of the doctoral program has taken shape over the course of seven years. For nearly a decade, he has been one of my most trusted advisors as I advanced my career, and he was the first to encourage me to apply to the Ph.D. program. Dr. Folse was by my side the very first time I presented at a conference. He was there again as a co-author and mentor on my first peer-reviewed publication, and he continued to share his wisdom during the dissertation process.

I would also like to thank my Co-chair, Dr. Joyce Nutta. Dr. Nutta's support and thoughtful consideration of my study was invaluable. I am incredibly grateful for the numerous afternoons spent at a coffee shop unpacking, deconstructing, and reassembling the concepts in my study and how they fit into the larger scope of TESOL and SLA. I would also like to thank Dr. Nutta for using this process to push my own academic writing skills by encouraging me to not aim for academic writing norms but to strive for eloquence, which is a quality that seems so effortlessly incorporated into her own writing. My success as a doctoral candidate is a direct result of the careful, steadfast mentoring of Dr. Nutta. Her ability to sense when I needed time to

reflect, protection from myself, words of encouragement, tough love, a push, or simply an ear to bend is uncanny.

Dr. Folse and Dr. Nutta's generosity and dedication to their students knows no limits. I am humbled and honored to have worked with them, and their example of academic integrity, professionalism, and leadership will continue to shape my professional and personal identify as I leave UCF and begin the next chapter of my life.

I would also like to thank Dr. Lihua Xu, my methodologist, for her patient and reliable guidance during this process. Even before agreeing to be on my committee, Dr. Xu answered numerous questions and helped me when I was in the weeds with my study design. Her guidance on my dissertation has developed my statistical knowledge and confidence.

I also owe a huge debt of gratitude to my fourth committee member, Dr. Kerry Purmensky, for taking on this project despite an already heavy thesis and dissertation advising load. Her thoughtful comments, flexibility, and encouragement during this process have been a great source of motivation. For years I have admired her as scholar and educator, so I was thrilled to have her on my committee.

I would also like to take this opportunity to thank other members of my UCF family. I would like to thank Dr. M. H. Clark for her help in the CASTLE lab. A huge thanks also goes out to the graduate students who gave up their time to participate in my study. I could not have finished without them! Finally, I would like to thank my cohort- Marcella Farina, Virginia Ludwig, Aimee Schoonmaker, Elizabeth Giltner, Donita Grissom, and Melanie Gonzalez. I am honored to be part of such an amazing group of strong, intelligent, accomplished, and hilarious women!

I must extend a special thanks to Melanie Gonzalez. As we discussed our applications to the Ph.D. program, we had no idea what we would be getting ourselves into, which was probably a good thing! I simply could not imagine going through this process without you. You are my comic relief, my cheerleader, and my co-conspirator! I look forward to many more adventures as we continue on our academic careers.

I would like to thank all the IEP directors, administrators, and teachers who participated in my study. I am incredibly grateful for your enthusiasm and interest in my topic. Without the 107 teachers that gave up an hour of their planning time, this study would simply not be possible.

Last, but certainly not least, I would like to thank my friends and family for their unwavering support during this journey. I would like to thank Dodie Ohrt, John Super, Darrell Murray, and Leigh DeLorenzi for offering their wisdom on how to survive this process! I would also like to thank Carmen and Ryan Hall, Jennifer Peppers, Erica and Dave Lucas, Liz Sanchez, and Angela Alexander for checking on me to make sure I was still alive, for putting up with multiple cancelled plans, for still claiming me when I brought a stats book to the beach or on vacation, and for showing a genuine interest in my research.

I must conclude this section with a humble word of thanks to my family. They did not make the choice to go back to grad school but nevertheless made sacrifices, and for this I am truly grateful. To my mom and dad, thank you for your support over the last 32 years. Thank you for instilling me with the dedication and drive to follow my dreams. To my sisters, Julie and Marilynn, and my brother-in-law, Matt, thank you for your cheers of encouragement and putting up with me. To my grandpa, George, thank you for reminding me to take a break and not work too hard. Thank you for your financial support during my data collection travels. Finally, thank

you to my grandma, Mary, who left this Earth two days before I began this journey. I miss you every day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

- AFL- Academic Formulas List
- EFL- English as a Foreign Language
- ESL- English as a Second Language
- GRE- Graduate Readiness Examination
- IRB- Institutional Review Board
- L1- Native Language
- L2-Non-Native Language
- MICUSP- Michigan Corpus of Upper Level Student Papers
- NNS- Non-Native Speaker
- NS- Native Speaker
- SAT- Scholastic Aptitude Test
- SPSS- Statistical Package for the Social Sciences
- TESOL- Teaching English to speakers of other languages
- WAC- Writing across the Curriculum
- WID- Writing in the Disciplines

CHAPTER 1: INTRODUCTION

It is hard to explain the university experience without discussing writing. Undergraduate and graduate students are tasked with producing literature reviews, lab reports, narratives, journals, blogs, and many other written artifacts to document their growing command of academia in general and their chosen field in particular. In fact, writing has become an important means of promoting and assessing student learning in post-secondary institutions thanks to the growth of the writing across the curriculum (WAC) movement.

The WAC movement is grounded in the work of Britton (1975) and Emig (1977) who advocated for writing as a means of learning. Britton (1975) noted that writing could function as a self-expression of knowledge as learners explored a topic. Journals, reflections, and other ungraded writing assignments could act as discovery learning (Mcleod, 1992). The second type of writing, according to Britton, is of a much higher-stakes variety.

Transactional writing is just as it sounds. A writer's composition is used as the currency of grades and achievement (Britton, 1975; Mcleod, 1992). Writing serves as a tool to demonstrate the mastery of a topic to others. It shifts the audience of the paper from the individual's learning to that of a larger social group. In transactional writing, WAC dovetails with writing in the disciplines, or WID. Writers are tasked with demonstrating content area knowledge and also begin an apprenticeship into the academic and disciplinary discourse communities (Mcleod, 1992). For example, in a literature review, students not only find, read, and synthesize studies on a particular topic, but incorporate discursive conventions from the readings into their own writing.

The International WAC/WID Mapping Project (2008) reports that 568 U.S. institutes of higher education have implemented WAC/WID programs, and the number is growing. Most of these programs were at large research universities that granted master and doctoral degrees. This finding is of particular importance in relation to second language learners.

Non-native English speakers are a growing demographic in U.S. institutes of higher education. Around 5 million non-native speakers will matriculate from U.S. public schools (Goldenberg, 2008; National Center for English Language Acquisition, 2011). Many of these students' were born in the U.S. Their parents are first-generation immigrants and a language other than English is spoken in the home (Ferris, 2009; Nutta, Mokhtari, & Strebel, 2012). They will most likely attend community college before transferring to a state university and may or may not pursue support from an English for Academic Purposes program (Ferris, 2009). Another source of linguistic diversity in university classrooms is international students. According to the 2013 Open Doors Report published by the Institute of International Education, there are 816,644 international students studying in the U.S. International students make up 4% of all university students, but these students are not equally distributed in universities across the country. In fact, 55% of international students are studying at large research institutions. An additional 37,000 are currently studying English at an Intensive English Program with hopes of pursuing a degree in the U.S.

The university classroom has become the point of intersection for writing initiatives in higher education and the challenges of second language writing development. As written assignments are the most common form of assessment of content mastery and course objectives (Ferris, 2009; Knodt, 2006; Sullivan, 2006; White, 2007), non-native speakers need to know how

to write and how to write well. Yet writing in a second language is challenging, and many second language learners remain ill-prepared for the linguistic demands of the university classroom despite years of preparation (Ferris, 2009).

First, academic writing requires academic language skills, which are slower to develop than social language and are pre-empted by access to the discourse community for input (Cummins, 1980; Gee, 1990). While simultaneously developing academic language skills, a learner must also differentiate between spoken and written academic discourse as conventions do not directly transfer from one form to the other (Biber, 1986; Biber, Conrad, & Reppen, 2004; DeVito, 1967; Halliday, 1979; McCarthy, 1998), and an adherence to spoken conventions in writing can mark a paper as amateurish (Cayer & Sacks, 1979). Academic text is also influenced by disciplinary conventions (Biber, 1986; Fang & Schleppegrell, 2010; Lee & Spratley, 2010), so while general academic writing skills are important, so are the nuances of a particular field of study. A second language student, even at an advanced level of proficiency, can face obstacles with language, mode, or discipline.

Before arriving in the university classroom, non-native speaking international students have to demonstrate their skills on language-specific exams such as the Test of English as a Foreign Language (TOEFL) and indirect measures such as the Graduate Readiness Exam (GRE). In order pass these tests, Intensive English Programs provide a necessary transitional learning experience, as many cannot reach required proficiency standards by studying in their home country alone. One of the key areas of an IEP curriculum is developing the writing skills of academically-oriented ELs so they can be successful on entrance exams and in the undergraduate

or graduate classroom. In order to accomplish this, a clear picture of what makes good writing is warranted, but this question provides as many slippery slopes as it does footholds.

The Council of Writing Program Administrators (2008) attempted to describe the baseline of writing skills for all university students, regardless of status as a native or non-native speaker. They outline skills in rhetoric, critical thinking and writing, process writing, and knowledge of conventions. Using these skills, student writers should be able to use a correct tone and level of formality and have an understanding of writing as a collaborative process within a community. Writers should be able to incorporate, evaluate, analyze, and synthesize primary and secondary sources clearly. Specifically, in the area of conventions, they should be sensitive to genre-specific approaches to writing structure, tone, and preferred grammatical structures. Sullivan (2006) presents a similar view of good writing. He emphasizes higher order thinking skills and strong discourse organization, while seamlessly weaving together abstract content, analyses, and sources. Adherence to grammar, punctuation, and spelling are also a must. Bloom (2006), however, argues that while scholars discuss what makes good writing, the fact is that students are not required to demonstrate good writing in the classroom, just writing that is good enough to blend in with "the lingua franca for writing throughout the writer's home institution…and meet the standard for writing beyond that college" (p. 71). Bloom cautions that readers "cannot afford to be distracted by departures from conventions of form, or language that calls attention to itself" (p.79). In other words, this mirroring effect, for better or worse, is what might actually determine acceptable writing.

In second language learning, native speakers have historically been held up as the standard L2 learners should try to emulate. Recently, the study of linguistic patterns used by

native and non-native speakers in a variety of contexts is gaining momentum thanks to developments in corpus linguistics. A corpus is a purposeful collection of language that identifies words, patterns of words, or even grammatical features that are used more frequently than by chance alone (Nation, 2001a). Corpora have been assembled to represent native speaker, non-native speaker, and a hybrid of both groups in particular disciplines. In writing, corpora analyses demonstrate prevalent patterns used by members of a particular discourse community and how far students are from meeting these community norms.

The current study focuses on corpus-identified patterns of words. The term used in second language acquisition and corpus linguistics to describe these patterns is formulaic language. Defining formulaic language, much like defining good writing, is no small feat. In fact, over 50 terms have been found in the literature to describe the phenomenon (Wray, 2002). One of the most basic definitions comes from Wray (2002) who describes:

> 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 (p. 9).

Researchers that have studied formulaic language from a phraseological perspective have narrowed this definition based on the transparency and substitutability of the sequence (Granger & Paquot, 2008; Howarth, 1998; Moon, 1998). Other researchers have operationalized salient patterns based solely on the frequency of occurrence (Biber et al., 1999; Cortes, 2004). Another group attempts to find the middle ground by relying on measures of frequency and cohesive meaning in order to identify subsets of language patterns that are not only frequent but pedagogically useful (Martinez & Schmitt, 2012; Simpson-Vlach & Ellis, 2010). While

researchers do not agree on the best approach to identify formulaic language, they do agree that it is abundant.

Conklin and Schmitt (2012) propose that it would be reasonable to attribute 30 to 50% of all language produced to formulaic sequences. This suggestion is congruent with several oft-cited studies. Biber et al. (1999) found that recurrent sequences made up 30% of conversation and 21% of academic writing. Howarth (1998) reported similar results in that between 30 and 40% of academic language consists of formulaic language. Altenburg (1998) reported the highest findings, as he attributed 80% of a sample of spoken English to frequent two-word combinations. Moon (1998) reported the lowest findings with less than 4% of a corpus of spoken and written English being covered by recurrent word sequences. The results largely depend on the operational definition and identification techniques.

Researchers suggest the reason that formulaic language is so common is connected to processing and production benefits of lexical patterns (Arnon & Snider, 2010; Pawley & Syder, 1983, Schmitt, Grandage, & Adolphs, 2004). In natural settings, Kuiper (1996) found that auctioneers and sportscasters relied heavily on formulaic language. Crystal (1995) also found highly formulaic language in weather forecasts. In experimental, cross-sectional designs, Millar (2011) found that departures from expected sequences of two-word collocations increased native and non-native speakers' reading and comprehension time. Ellis and Simpson-Vlach (2009) studied native and non-native speakers' ability to judge if a sequence of words was correct in fill-in-the-blank sentences. Both groups had a faster judgment response time when the blank was filled with a sequence, but native speakers were faster at processing sequences that had a stronger measure of cohesiveness between the words. Non-native speakers responded faster to

sequences that were more frequent. Schmitt et al. (2004) also found that processing and recall time in a dictation task were faster and more fluent when formulaic language was incorporated. However, not all sequences were reproduced with the same ease, and the researchers concluded that what is stored as formulaic might be unique to each learner and the input they received. While the research does support benefits in language processing, it is unclear how these benefits might be different between native and non-native speakers.

Even though formulaic language influenced the processing for both groups of speakers, there is a large gap between native and non-native speakers' use of sequences as evidenced in writing. Yorio (1989) and O'Donnell, Römer, and Ellis (2013) found that non-native speakers consistently used less formulaic language than their native speaking peers, and advanced non-native speakers used more sequences than students with a lower proficiency level. Howarth (1998) looked at the writing of native and non-native speaking graduate students and found a gap of over 50% in the use of sequences and idioms. Granger (1998), Hewings and Hewings (2002) and Li and Schmitt (2009) found that non-native speakers overall used less sequences, but the overuse of a specific subset of sequences was the most damaging in marking writing as developing or substandard. Furthermore, attempts at direct instruction to close the gap have produced lack-luster results (Jones & Haywood, 2004; Schmitt, Dornyei, Adolphs, & Durow, 2004).

Rationale

Evidence suggests that formulaic language is beneficial in language processing and L1 and L2 speakers use different amounts of vocabulary strings, yet little is known about the practical effect of this gap between native and non-native language. Based on the research of

single-item vocabulary words and perceptions of second language writing, this gap may be important. Engber (1995) found a significant relationship between IEP instructors' grades on a timed writing assignment and the vocabulary in the writing assignment. Essays that used more words only once and included less lexical errors scored higher. Santos (1988) found that professors were more accepting of grammatical errors in second language writing, but marked cases with the wrong vocabulary word as the most irritating and distracting. Furneaux, Paran, and Fairfax (2007) found that EFL teachers from five countries almost exclusively focused on vocabulary when providing feedback on students' effectiveness as a communicator. Second language writers' ability to demonstrate vocabulary knowledge in a writing assignment does affect teachers' perceptions of quality, or in other words grades.

Research on what it means to know a word also clearly extends knowledge beyond the white space on either side of a group of letters. Knowing the context and collocations of a particular word are an integral part of vocabulary knowledge (Folse, 2004; Nation, 2001b). Schmitt (1998; 2010) notes that this level of vocabulary knowledge may take longer to acquire and is reflective of an advanced mastery of the L2 lexis, which supports the findings that use of formulaic language increases as a function of proficiency and experience (Li & Schmitt, 2009; O'Donnell et al., 2013; Yorio, 1989).

In the case of single-item vocabulary words, corpus-based lists are used to help second language writers improve their knowledge and command of vocabulary in academic settings (Nation, 2001a). Vocabulary lists such as the General Service List (West, 1953), and the Academic Word List (2000) are used to make decisions about what words to teach, evaluate the vocabulary difficulty of a text, or determine the complexity of a student's writings. If vocabulary

knowledge does extend to include the company a word keeps (Firth, 1957) and a learners' ability to demonstrate vocabulary knowledge in writing does influence score, then it is reasonable to test if formulaic language influences teachers' perceptions of quality, as measured by writing score.

In fact, only one study to date has directly investigated the effect of formulaic language on writing quality. Ohlrogge (2009) studied the relationship between scores on an intermediate-level timed writing proficiency test and six types of formulaic language: idioms, collocations, transitions, phrasal verbs, personal stance markers, and generic rhetorical phrases. He found that idioms, collocations, and personal stance markers were associated with higher ratings on a five-level holistic rubric. However, as the study looked at a wide range of formulaic language, it is unclear how to replicate and apply these results to the second language classroom. Also, it is unclear how much formulaic language is needed to affect instructors' perceptions of writing.

Few formulaic language studies translate their results into pedagogically applicable forms. Only two corpus-based lists have been generated- the PHRASE list (Martinez & Schmitt, 2013) and the Academic Formulas List (Simpson-Vlach & Ellis, 2010). These corpus-based lists have yet to be applied in studies that measure the effect of formulaic language on L2 language output or L2 classrooms.

Classic guides to composition such as Vrooman (1967) and Strunk and White (2000) allude to the benefits and drawbacks of lexical patterns in writing. Vrooman cautions new writers to "abandon well-worn phrases that come easily to mind and find words that will give individuality and freshness to your…ideas" (p. 64) while at the same time suggesting it is equally detrimental to "strain for novelty" (p. 64). Strunk and White suggest beginning writers "err on the side of conservatism, on the side of established usage" (p. 83) in the words they use. While

not directly referring to corpus-identified formulaic language, these guides demonstrate a well-established pull between novelty and adherence to norms in writing. In all, it is unclear how formulaic language fits into the discussion of effective writing and the role is should play in preparing second language learners to be successful in higher education classrooms and ultimately their chosen professions.

Research Questions

Given the demands placed on ESL teachers to prepare students for academic settings and the prominent use of writing as a form of assessment in higher education, the role of vocabulary in successful second language writing is an important area of interest. More specifically, there is a growing body of theoretical and empirical studies suggesting that vocabulary knowledge and production should expand to include formulaic patterns of language and that second language learners greatly differ in their pattern-based lexical knowledge. What is not clear in the research is what formulaic patterns would be beneficial to learners if they were taught in ESL and EFL classrooms. Therefore, the aim of the current study is two-fold. While the specific research questions and explanation of variables and instruments are included in Chapter Three, the following paragraph outlines the main queries of the study.

First, the primary question in this study is to evaluate the effect of formulaic language on scores given by ESL writing teachers on academic writing assignments. Second, the study also looks at the number of formulaic sequences to determine if the amount can mitigate the scores given by ESL teachers. Finally, the study investigates any potential differences in teacher perceptions reflected by scores on writing skill-specific sub-scales of a grading instrument.

Potential Contributions of the Study

The current study is based on the findings that non-native speakers, even at the advanced level, use less formulaic language than their native-speaking peers (Durrant & Schmitt, 2009; Granger, 1998; Hewings & Hewings, 2002; Li & Schmitt, 2009; O'Donnell et al., 2013) and tend to overuse a set group of formulaic sequences (Granger, 1998; Hewings & Hewings, 2002; Howarth, 1998) which serves to categorize their writing as novice by seasoned members of the discourse community (Altenberg, 1998; Cortes; 2004; Granger, 1998; Moon, 1998; Wray, 2002). Therefore, the presumption is that by increasing the amount of formulaic language, a writer can increase his or her score on a writing assessment.

The study makes a significant addition to the body of research because it fills a gap in the literature between the presence, teachability, and the value of formulaic language. Researchers, as mentioned above, have documented the uneven distribution of formulaic language in second language writing. Also, the few studies on the learnability of formulaic language have not shown promising results (Cortes, 2004; Schmitt, Dornyei, Adolphs, & Durow, 2004; Jones & Haywood, 2004). However, only one study to date has looked at the advantage of formulaic language and writing assignment score, which are a reflection of teacher perceptions of quality. The study found that various types of formulaic language increased in frequency as a function of writing assignment score (Ohlrogge, 2009). If formulaic language is found to influence writing score, then more research on effective pedagogical approaches is needed.

Second, in the current study the grading instrument used to generate the writing score is out of a possible 100 points and includes five sub-scales representing higher order and lower order rhetorical skills. By looking at a rubric with a range of possible scores and five sub-scales,

the study adds to the understanding of how formulaic language is perceived. Formulaic sequences have lexical, discursive, and grammatical functions (Halliday, 2004; Simpson-Vlach & Ellis, 2010; Wray, 2002). If evaluating changes in composite score alone, valuable information could be lost. For example, two students might have the same score on a writing assignment but have very different competencies in organization, grammar, or vocabulary. The same logic applies to formulaic language and writing score. A non-significant difference in composite score may hide significant changes in discrete skills assessed implicitly. Understanding where a change in composite score might or might not occur contributes to the research on the lexiogrammatical (Halliday, 2004) characteristics of formulaic language.

Furthermore, by studying the possible score changes in the sub-scales, the research adds to the field's understanding of vocabulary instruction. Currently, vocabulary instruction is often relegated to the reading teacher even though it is important across the curriculum (Folse, 2010). Formulaic language is essentially an extended collocation, which Folse (2004) and Nation (2001b) established as a key element in vocabulary knowledge. Since previous research has been unable to entirely operationalize these sequences as enhancing either lexical, syntactical, or discursive properties of a writing assignment, this study sheds light on the potential benefits and drawbacks of their instruction in writing, reading, grammar, and other courses.

Definition of Terms Used in the Study

- *Advanced second language learner*- A student whose native language is not English and is enrolled as a degree-seeking student at a post-secondary institution that uses English as a medium of instruction

- *Compositionality*- A term used in phraseological studies of vocabulary that describes how resilient a phrase would be if any of the key words were replaced with a synonym
- *English as a foreign language (EFL)*- Describes an environment where the students are studying the target language, but it is not the dominant language of their surroundings
- *English as a second language (ESL)*- Describes an environment where students are studying and living in an environment where the target language is the dominant language
- *Formulaic sequence*- A series of three words taken from the Academic Formulas List (Simpson-Vlach & Ellis, 2010) that occurs together more often than by chance in academic writing of native speakers, operates as a cohesive group, and is found to be pedagogically salient by English language teachers
- *Intensive English Program (IEP*)- An institute that prepares adult ESL students to study a variety of fields at a university where English is the medium of instruction
- *Lexical density*- The ratio of content words to total running words
- *Native or first language (L1)*- The language learned from birth
- *Native speaker (NS)*- An individual operating in an environment that uses the language they have acquired from birth; or a monolingual speaker
- *Non-native speaker (NNS*)- An individual using a language that they acquired as a second, third, or additional language
- *Second language (L2)*- The additional language learned as a child or as an adult
- *Transparency*- A term used in phraseological studies of vocabulary that describes how easily the meaning of a group of words can be deciphered from the individual parts

- *T-Unit*- A measure of writing complexity that divides discourse into the smallest grammatically correct complete unit of meaning (Hunt, 1964)
- *Type-token ration*- The ratio of unique words used only once to total running words

Organization of the Study

Chapter One contextualized the current study within the present state of writing instruction for academically-oriented language learners. The chapter also presented an overview of the research questions followed by a discussion of the study's potential contributions. Finally, a list of key terms used in the study were explained.

Chapter Two grounds the study in the current literature of the field. The chapter discusses the nature of academic language, the role of vocabulary in academic language proficiency and vocabulary size requirements, and then presents an argument to expand the notion of vocabulary to formulaic sequences. The qualities of formulaic language and their potential contributions to second language acquisition are also reviewed. Finally, an overview of the role of formulaic language in academic writing is presented.

Chapter Three introduces the methods to be used to isolate formulaic language, create academic writing samples, and collect data from participating ESL writing teachers. The instruments used to collect data are explained and empirical support for their validity is presented. Finally, the statistical analysis used in the study are reviewed.

Chapter Four includes a description of the non-native writing samples used to collect data. Steps taken to clean the data and check for statistical assumptions are outlined. The chapter ends with the results from a series of analyses to answer the research questions.

Chapter Five reviews the rationale of the study and the findings from Chapter Four. The final chapter then presents a discussion on the significance of the findings, limitations of the study, and suggestions for future research. Chapter Five concludes with implications for the classroom.

CHAPTER 2: RESEARCH AND LITERATURE REVIEW

More than three decades ago, Meara (1980) argued that vocabulary acquisition was a neglected area in the field of second language acquisition (SLA). Meara pointed out that syntactic, morphological, and phonological characteristics of language learning received the most attention from researchers, teachers, and materials writers. This focus is no longer the case as the role of lexical competence in language proficiency is well established, and researchers have investigated the quantity of vocabulary words a learner needs to acquire in order to successfully communicate (Cummins, 1982; Hu & Nation, 2000; Nation, 2006; Schmitt, 2010). Other researchers have focused on identifying what words should be a part of a learner's lexical repertoire through analyzing purposeful samples of language-in-use in general (West, 1953; Browne, Culligan, and Phillips, 2013; Brezina & Gablasova, 2013) and academic (Coxhead, 2000; Gardner & Davies, 2013) environments. Other scholarly groups have called for a discipline-by-discipline survey of the lexis (Hyland & Tse, 2007; 2009). In fact, the attention toward vocabulary-related research suggests that the field has adopted the view that "While without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (Wilkins, 1972, p. 111).

Regardless of the area of vocabulary interest, changes in technology and computer programming have dramatically altered the manner in which the second language (L2) lexis is studied. Corpus-based research has become more and more accessible for researchers, materials writers, teachers, and even students. A corpus, or a purposeful sample of language, allows researchers to look at how language is used by native speakers as opposed to relying on intuition about what sounds natural, which can be unreliable (Folse, 2004; Nation, 2001a; Sinclair, 1991).

Yet, an important contribution of these language analyses is not related to specific individual words per se, but in uncovering pre-packaged groups of words that occur together more often than by sheer coincidence.

These formulaic sequences have pragmatic as well as discursive functions. They are an insight into the intersection of vocabulary with other language subsystems such as grammar and show how words help unite a speech community. Formulaic sequences are extremely common in both spoken (Biber, Johansson, Leech, Conrad, and Finegan, 1999; Biber, Conrad, and Cortes, 2004; Crossley and Salsbury, 2011) and written (Biber et al., 1999; Biber et al., 2004; Cortes, 2004; Howarth, 1998) contexts. Research also shows that second language learners are cognizant of the language bundles that frequently occur in the input they receive (Durrant & Schmitt, 2009; Ellis & Simpson-Vlach, 2009; Ellis, Simpson-Vlach, & Maynard, 2008). Yet, when studying adult second language writing, there appears to be a significant gap between the number and type of sequences used when compared to their native-speaking counterparts (Durrant & Schmitt, 2009; Granger, 1998; Hewings & Hewings, 2002; Li & Schmitt, 2009; O'Donnell et al., 2013). To compound these findings, initial studies on the durability of formulaic sequences after direct instruction show marginal effects if any (Cortes 2004; Jones & Haywood 2004; Nesselhauf & Tschichold 2002; Schmitt et al., 2004). Therefore, research is needed to understand the possible connection between formulaic language and writing performance. If the integration of formulaic sequences into a writing assignment can increase a teacher's perception of writing performance, i.e. the score, then it will add weight to the descriptive research on the productive gap between native and non-native speakers and support the role of formulaic language in language

processing. Furthermore, it will ground the research on pedagogical interventions to promote the acquisition and use of formulaic sequences.

This chapter begins with a review of the literature on the nature of language proficiency in general followed by connections to vocabulary knowledge and academic writing. Next, an overview of corpus-based single-item vocabulary lists is presented. The discussion then focuses on theoretical support for shifting away from single-item vocabulary word lists to multiword sequences. The role of formulaic patterns in second language acquisition and language processing is then presented. Finally, corpus-based literature on the prevalence of formulaic language in writing, the gap between native and non-native speaker use of formulaic language, and the growth of formulaic language knowledge as a result of instruction are reviewed. The chapter concludes with a discussion of pedagogically-minded corpus-based lists of academic formulaic sequences.

The Nature of Language and Vocabulary in Second Language Acquisition

The concept of language proficiency is often presented anecdotally as a dichotomy. One is either 'fluent' or 'not fluent', but in fact it is a multi-level process extending beyond the relationship of knowledge between pronunciation, morphology, syntax, and vocabulary. Cummins (1979; 1981) first posited a hierarchy in language proficiency by classifying language as Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). BICS is the language of day to day social interaction and is typically acquired within one to two years of entering the target language environment. CALP, which takes between five to seven years to acquire, represents an ability to communicate in a school setting in both oral and written contexts. To illustrate the importance of this distinction,

Cummins (1980) evaluated 400 referrals to school psychologists of English learners in the Canadian public school system. Cummins found that the referrals stemmed from teachers' and administrators' confusion by students who could communicate freely with them but performed far below expectations on classroom work and formal assessments. He found that, in some cases, the student's status as a language learner was not even included in a referral requesting a remediation plan.

Other researchers propose similar distinctions in language proficiency. Gibbons (1991) mirrored the BICS and CALP distinction in her discussion of social and classroom language. She argued that social language:

> Does not normally require higher order thinking skills such as hypothesizing, evaluating, inferring, generalizing, predicting, or classifying. Yet these are the language functions which are related to learning and the development of cognition; they occur in all areas of the curriculum, and without them a child's potential in academic areas cannot be realized (p. 3).

Gee (1990) further supports the hierarchy of language proficiency by using the terms primary and secondary discourse. According to Gee, primary discourse is learned face to face, at home, and through social interaction. Secondary discourse is acquired through some sort of social institution, for example a university, where specific features and usage of the language have been deemed appropriate for communication among its members. Cummins (2007) sums up the distinction between social and academic language as "the life chances of individuals are directly determined by the degree of expertise they acquire in understanding and using this language" (p.

7). The distinction between BICS and CALP, although a simple one, illustrates that language proficiency does not neatly fit into one domain.

Language proficiency is also not as simple as academic or social. In both cases, successful communication depends on multiple, interacting skills. Building on the communicative competence research of Canale and Swain (1981), Bachman (1990) presents two skills that feed into a learners' language capability. The first skill, organizational competence, relates to the application of grammatical rules to produce sentences and to further connect these utterances together into discourse with a cohesive meaning. The second skill, or pragmatic competence, deals with the effective application of language functions. A person must be able to use language to perform illocutionary tasks while simultaneously accounting for sociolinguistic factors such as politeness, dialect, and register to name a few. In considering Bachman's explanation of communicative competence, it is clear that communication requires simultaneous interactions between rule-based and meaning-based awareness. Later in this chapter, a theoretical discussion is presented that argues form and meaning do not necessarily represent mutually exclusive systems.

Aside from social and academic language, research indicates that there is also a difference between spoken and written language. CALP is most often produced and assessed in high stakes situations through reading (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Lervåg & Aukrust, 2010) and writing (Cohen, 2005; Engber, 1995; Santos, 1988). The lexical standards of appropriate communication differ between the two registers, and proficiency in speaking and listening do not always directly translate to reading and writing (Biber, 1986; Biber et al., 2004; Coxhead, 2000; Simpson-Vlach &Ellis, 2010).

Halliday (1979) states that spoken language is representative of complex sentence structures that contain fewer content words, or have a low lexical density. Written language, however, has denser vocabulary compacted into simpler sentence structures. DeVito (1967) compared the academic writing and speech patterns in an interview of ten university professors. DeVito's findings mirror Halliday, as he found written language to have more lexical diversity, make use of low frequency, content specific-words, use more nouns and adjectives than verbs and adverbs, and use simpler sentence structures overall.

A seminal work on the textual differences between spoken and written language comes from Biber's (1986) investigation of 41 linguistic features in 545 English language samples totaling 1 million words. Based on a review of the literature, Biber selected linguistic features that represented eight hypothesized distinctions between speaking and writing (p. 388):

1. Writing is less interactive and more detached through the use of passives and nominalizations
2. Writing uses more elaboration through subordinate clauses and prepositional phrases
3. Writing uses more precise vocabulary and has more lexical diversity
4. Writing more explicitly denotes the relationship between ideas
5. Speaking is more informal and less explicit due to informal vocabulary and reliance on the pronoun *it* as a reference
6. Speaking uses more first and second person pronouns and questions to create an interactive style
7. Speaking uses more place and time adverbs to show a focus on the environmental and temporal context

8. Speaking and writing use different verb tenses

Biber wrote a computer program to analyze the 41 features as evidenced by occurrences of 420 of the most common verbs from the Lancaster-Olso-Bergen Corpus. Biber then conducted a factor analysis using the data on patterns of verb occurrences. Biber only included factors that had a factor weight of at least .35 in the final analysis. In Biber's study, factor weight is a measure of the extent to which a variable is represented in the spoken or written language samples. Factor weights can be either positive or negative, and when looked at together can shed light on what does and does not co-occur. The results indicated three main clustering of factors which Biber termed interactive vs. edited, abstract vs. situated, and reported vs. immediate.

The first factor, interactive vs. edited, included positive factor weights for linguistic features such as hedging (.61), use of the pronoun *it* (.49), yes-no questions (.79) and wh-questions (.52). Negative factor weights were only found for word length (-.71) and lexical diversity (-.65). Biber suggests these factors represent the tendency of written language to maximize the content in the fewest words possible which requires explicit vocabulary choices. Furthermore, he argues that due to the time constraints of oral communication, speakers compensate by using complex sentences as opposed to complex vocabulary.

The abstract vs. situated factor includes positive weights for the use of nominalizations (.74), prepositions (.61), passives (.47), and word length (.40). Linguistics features with negative weights include 3rd person pronouns (-.25), adverbs of time (-.55) and place (-.57), and deletion of the relative pronoun (-.50) and the subordinate *that* (-.42). Biber describes this combination of characteristics as representative of writing as a learned process. Consider the example *the*

experiment was carried out under normal conditions. The use of the passive voice is explicitly taught as a function of academic writing. Biber suggests factor two points again towards writing's tendency to compact information to increase the efficiency of each word. The reported vs immediate factor contrasts the use of features such as past tense (.89) and 3rd person pronouns (.61) with present tense (-.62) and adjectives (-.40). Here, Biber suggests the positive weights are representative of a narrative style that recounts information compared to active, in-the-moment communication.

Biber then calculated the mean factor score for text types in the corpus and plotted them on a factor-level graph to show where each category of spoken and written language fell on a spectrum of characteristics (as illustrated in Figure 1). What Biber's research ultimately shows is that spoken and written language have both similarities and differences on a multi-dimensional level. For example, academic prose and news broadcasts are similar in terms of factor one but almost completely opposite in terms of factor two. In other words, written and spoken language in these scenarios are linguistically similar in their limitations on interaction with the audience and use of lexical diversity. They are linguistically different, however, in the level of compression of information into sentences and the degree of reference to the current time and place.

In sum, writing and speaking are certainly unique systems of communication that have evolved to include unique standards for proficiency in each (Allen, 1966; Chafe & Tannen, 1987; Pawley & Syder, 1983); nevertheless, the differences are not unilateral (Biber, 1986). These differences, however marked or subtle, are important for academic-oriented language learners. Cayer and Sacks (1979) studied college freshman (*n=8)* in remedial writing and reading courses.

The researchers asked participants to read a controversial statement, discuss it, and then write an essay about it. They noted similarities in their oral and written discourse; concluding that their reliance on spoken language norms in writing contributed to their remedial classification.

Further anecdotal support for the distinction between written and spoken language comes from a standard methodological step in corpus-based studies. In researching academic writing, researchers evaluate the applicability of their corpus findings in non-academic writing or speech samples (Coxhead, 2000; Simpson-Vlach & Ellis, 2010). A lack of support in the opposing corpus is used to validate the credibility of the findings to the original communicative context.

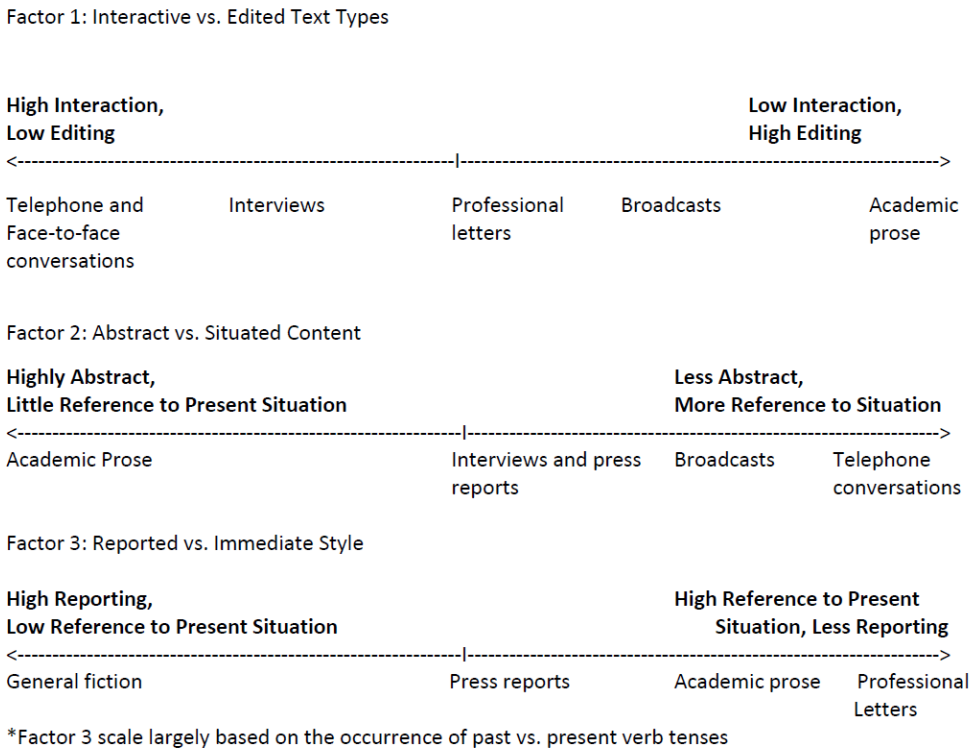


Figure 1. Biber's (1986) Linguistic Features of Speaking and Writing

To further complicate matters, some researchers suggest that proficiency is dependent on a third axis, or discipline-specific proficiency standards. In comparisons of written texts alone, research suggests that successful communication differs greatly between disciplines in terms of vocabulary, preferred structures, and the tolerance for abstractness (Fang & Schleppegrell, 2010; Graesser, McNamara, Louwerse, & Cai, 2004; Lee & Spratley, 2010; Shanahan & Shanahan, 2008), which connects back to Biber's (1986) conclusion that not all writing is different in the same way. In terms of second language (L2) learners, Hyland and Tse (2007; 2009), are adamant that a one size fits all academic vocabulary is not helpful and cite an uneven distribution of corpus-identified general purpose vocabulary words across disciplines like the humanities and hard sciences. A more in-depth discussion of the disciplinary criticism of general purpose academic vocabulary can be found in the section on corpus-based vocabulary lists.

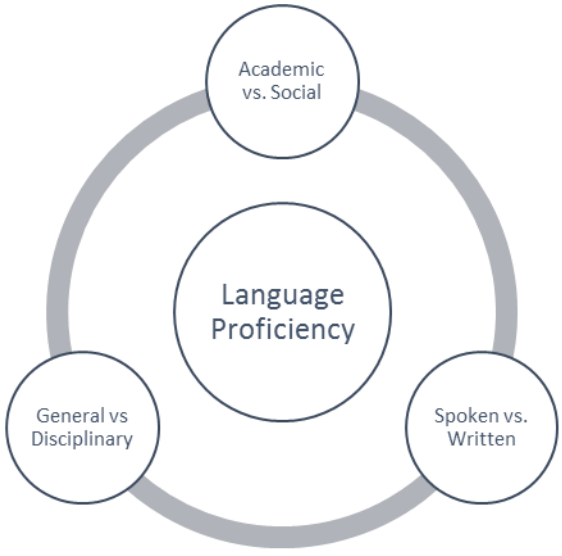


Figure 2. A Visual Model of Language Proficiency

Evidence of Vocabulary as the Cornerstone of Proficiency

The body of vocabulary-focused empirical research has expanded, and as a result, there is a clearer picture of how many and what kinds of words should be in the lexis of English learners. In terms of quantity, the suggested vocabulary size can vary depending on the scenario (general vs. academic) and the assumed threshold for unknown words without affecting communication. Two commonly referenced thresholds (Hu & Nation, 2000; Nation, 2006; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, 2008) are 95% and 98%. At the 95% threshold, one in every 20 words would be unknown, or in a written text, this would average to one unfamiliar word every two lines of text. At 98%, this threshold drops to only one unfamiliar word out of 50, or every five lines of text (Nation, 2006). Schmitt (2010) estimates the vocabulary size for speaking to be around 2,000 to 3,000 word families at the 95% level, or 6,000 to 7,000 at 98%. In reading or writing, the estimates of vocabulary size increase. Laufer and Ravenhorst-Kalovski (2010) suggest the lowest figure at 4,000-5,000 word families while Nation (2006) suggests 8,000 to 9,000. These numbers alone are intimidating, but they actually represent exponentially larger figures because the research uses word families as the standard unit of counting (*favor, favors, favored, favorite, favorable, unfavorable*).

Aside from the large quantity of vocabulary needed, the concept of knowing a word is also multidimensional. Nation (2001b) provides three key areas of knowing a word: form, meaning, and usage. Form includes the aural and visual representations. For example, the words *colonel* or *receipt* do not look the way they sound, and in fact the connection between the forms need to be explicitly taught (Folse, 2004). Folse (2004) and Laufer (1990) also point out that generating the parts of speech require the intersection of form and meaning. Understanding what

morphemes to use and recognizing irregular orthographic shifts can determine how easy a word is to learn and use correctly. Words can also have multiple meanings, and in fact the most common words in English are polysemous (Crossley, Salsbury, & McNamara, 2010; Folse, 2004). The word *class* can mean a group of students studying together, a social standing in society, a biological categorization, or refined behavior. Not all meanings are equally common in use and will not provide equal growth in language skills (Nation, 2001a).

Word meaning is also influenced by connotation which provides subtext as in *thin, slender, lean,* and *skinny* (Folse, 2004, p. 11). *Slender* and *lean* are more complimentary than the other synonyms. Meaning can also be influenced by the learner's native language. Cognates, or words that appear to be semantically related, can be a useful tool for vocabulary learning such as *traîne* in French and *train* in English. Other words can have false relationships, such as *embarazada,* or pregnant, in Spanish and the English word *embarrassed*. After form and meaning, the appropriate use of the new word, including collocations and register (Folse, 2004; Nation, 2001b), are important. This knowledge describes the environment of a word at the phrase, sentence, and context-level.

In sum, learning any new vocabulary word simply cannot be completed in one classroom sitting. Of all these facets of word knowledge, some are learned simultaneously while others take repeated instruction, practice, and input to master (Schmitt, 1998). Schmitt (2010) notes two ends of the vocabulary learning spectrum. The form-meaning association is an initial reflection of word knowledge while collocation and register reflect more advanced skills. The quantity of vocabulary knowledge combined with the multiple levels of knowledge all contribute to what Laufer (1997) describes as the 'lexical plight' of second language acquisition.

Categorizing an English Learner's Vocabulary

Words are typically categorized as either content or function words. Content words are namely nouns, verbs, adjectives, and adverbs. Function words such as prepositions, auxiliary verbs, and pronouns add grammatical information. In relation to second language learning, Nation (2001b), separates the L2 vocabulary into four categories: high frequency, academic, technical, and low frequency. High frequency words can be function words such as conjunctions (*because*) or prepositions (*at, by*) and common content words (*school, develop*). These words make up around 80% of spoken and written language (Nation, 2001b; West, 1953). Academic words are those often found in textbooks, journal articles, and university lectures but not common to daily social interactions (*imply, perspective*). Coxhead (2000) found that academic vocabulary accounts for approximately 10% of words in scholarly settings. Next, technical words are not common in social exchanges or academia at large, but rather in specific disciplines (*scalpel, incision*). Nation (2001b) estimates that technical words account for around 5% of words in any given exchange. Finally, low frequency words, despite the name, actually represent the largest group of words in the language as a whole. They are difficult to classify because it is dependent on the language exposure for any particular student. In short, one student's low frequency words could be another's technical words (Xu & Nation, 1984).

These categories are helpful in discussing the amount of vocabulary needed, but the taxonomy is not without weaknesses. Nation (2001b) points out that there is no clear dividing line between these categories, especially in terms of 'high' and 'low' frequency. Furthermore, the distinction of 'academic' and 'technical' vocabulary connects to the debate of general vs.

disciplinary academic proficiency mentioned previously, and it has direct implications for corpus-based studies.

Evidence of Vocabulary as the Cornerstone of Effective Writing

While the previous section presented a macro view of vocabulary in second language acquisition, Engber (1995) aptly points out "Vocabulary is not usually learned for its own sake. An important aim of a vocabulary program is to bring learners' vocabulary knowledge into communicative use" (p. 309). Vocabulary is a predictor of successful second language writing (Guo et al., 2013; Engber, 1995; Gonzalez, 2013; Grobe, 1981; Laufer & Nation, 1995). Guo et al. (2013) studied the relationship between 21 linguistic variables and raters' scores on a timed independent writing task for the TOEFL iBT (*n=240*). The resulting model was significant, *F(1, 154) =57.33, p<.001, r=.81*, for five linguistic features and accounted for 65% of the variance in score. Of these five features, three were vocabulary-related; contributing 62% of the explained variance. The number of words in the essay (47.8%), the average syllables per word (9%), and the number of noun hypernyms (4.8%) all significantly contributed to the model. To clarify, hypernyms are commonly referred to as umbrella terms. Dogs, cats, birds, and fish could all be classified under the general noun pets (Guo et al., 2013). The study shows that elements of vocabulary, not grammar, were the greatest determiner of essay score. It should be noted that Guo, Crossley, and McNamara did not triangulate the relationship between vocabulary, grammar, and errors. To discuss the role of errors, it is important to consult the work of Engber and Santos.

Engber (1995) studied the extent to which lexical density, lexical error, and lexical variation affected perceptions of composition quality for non-native speakers at an intensive

English program. Lexical density is the ratio of content words (nouns, verbs, adjectives, adverbs) to total number of running words. Lexical error is the ratio of form or meaning-based vocabulary errors to total number of running words. Lexical variation includes the ratio of unique, non-repeated vocabulary words to total running words. Engber asked IEP instructors (*n=10)* to rate 66 timed writing samples on a six-point holistic rubric of writing quality. Using a Pearson correlation, Engber found no significant relationship between score and lexical density (*r=.23, p>.01*); there was, however, a significant correlation between lexical error and score (*r=-.43, p<.01)*. As the number of lexical errors decreased, the score increased. Lexical variation also significantly correlated with score regardless if lexical errors were included (*r=.45, p<.01*) or excluded (*r=.57, p<.01*) from the calculation. Engber concludes that raters' perceptions of NNS writing is affected positively by the variety of vocabulary words in an essay and negatively by errors in word form and usage.

Santos (1988) evaluated professors' (*n=178)* ratings of content and language in NNS writing. Professors graded one of two 400-word compositions on a rubric with six subscales related to content impression, development, and sophistication. The other three subscales evaluated language comprehensibility, acceptability, and irritation. Participants first rated an essay with all original errors produced by the non-native writer. Based on the results of the first round of ratings, the researcher corrected all but the most salient errors. The remaining errors were underlined, and participants rated the sentence that included the error on only language comprehensibility, acceptability, and irritation.

Using Duncan's Multiple Range Test, Santos found a statistically significant difference in professors' harshness of grading content (*M=5.02, 5.76)* compared to language (*M=6.25, 6.79)*

for both NNS essays. From the second sentence-level rating, Santos looked at the overall rank order of the most salient errors in both essays. Sentences that included lexical errors were ranked as the most severe. Santos concluded that professors can and do distinguish between a NNS' content and linguistic knowledge and judge these two elements separately. However, Santos found that lexical errors are the exception, as "it is precisely with this type of error that language impinges directly on content; when the wrong word is used, the meaning is very likely obscured" (p. 85). The ability to overlook the developing language of English learners in writing is less likely to occur at the lexical compared to the grammatical level.

In a later qualitative study on instructor feedback on written assignments, Furneaux, Paran, and Fairfax (2007) looked at the written corrective feedback of EFL teachers (n=110) from five countries. The researchers coded the written feedback based on the purpose and the point of view of the teachers' comments. In the latter coding scheme, the researchers were looking for information on teachers' grading stance as either a fellow interlocutor or as a gatekeeper of knowledge. The researchers found that teachers mainly focused on grammar from a gatekeeper stance. However, feedback coded in the communicative category focused almost exclusively on vocabulary.

Second language learners are also aware of the importance of vocabulary in writing. Leki and Carson (1994) surveyed non-native speakers (*n=128)* at the end of their writing course in an English for academic purposes program. Students overwhelmingly reported a desire for more vocabulary instruction. Folse (2010) found similar results as students rated instructors that focused on vocabulary higher, regardless of the subject or skill area. Coxhead (2012) investigated second language learners' perceptions of vocabulary in writing at the post-

secondary level. Participants (*n=14*) were ESL students in New Zealand who read a 400 word text and wrote an essay based on the reading. In a semi-structured interview protocol, she found that participants reported a "need for technical, academic, or professional words to express ideas in writing" (p. 139). Participants were actively aware of their use of academic vocabulary during the writing process. For example, subjects added and deleted words in hopes of making their paper sound more academic and better received by an academic audience. In a qualitative study by Gonzalez, Youngblood, and Giltner (2012) in a beginning French as foreign language classroom, the researchers found that 42% of student-initiated requests for help during an in-class writing session focused on vocabulary.

These studies demonstrate that vocabulary is in the forefront of both teachers' and students' minds in the second language writing classroom. Much like a house, the foundation and frame of a message are made up of words. While vocabulary is not the only element of successful second language writing, it is a fundamental element.

Corpus-Based Single-Item Vocabulary Lists

Teachers, material writers, and students look to researchers to funnel and apply vocabulary research into pedagogically-appropriate tools. A main avenue through which scholars expand the knowledge-base and aid instruction is with corpus-based research. Corpus studies identify words that will give learners the biggest gains in proficiency as not all items increase decoding power to the same degree (Nation, 2001a). A corpus is a purposeful selection of language that, when analyzed, can reveal information about how communities of native speakers use words. There are four important single-item word lists that directly relate to the present study.

Defining a Corpus and Corpus Methodology

As a corpus-based word list is only as good as the corpus it comes from, there are several guidelines in evaluating corpora. In general, a good corpus:

a) Contains a minimum of 1-3 million words when looking for high frequency, general vocabulary words (Brysbaert & New, 2009; Coxhead, 2000). Corpora with less than a million words are appropriate for qualitative research (Granger, 1998)

b) Includes language samples that are generalizable to the linguistic environment of the intended learners and closely matched to the research question (Biber et al., 1999; Davies, 2010; Gardner & Davies, 2013; Nation & Webb, 2010). In corpus-based studies, words replace people. The concepts behind sampling procedures that maximize generalizability must be followed.

c) Includes a variety of authors at different lengths of texts to avoid bias from one particular writing style (Biber et al., 1999; Coxhead, 2000; Wray, 2002)

After assembling a sample of language that meets these criteria, a combination of a priori characteristics selected by the researcher identifies the 'most important' words. One of the fundamental characteristics of corpus-based research is the use of frequency, or the number of times a word occurs in the corpus.

There are two main counting units of word frequency: the word family and the lemma. A lemma is a smaller unit of counting as it represents a headword such as *develop* and its inflectional, or grammatical, variants such as *develops, developed*, and *developing*. A word family includes all the forms in a lemma plus derivations like *development, undeveloped, underdeveloped, developer*, and even inflectional variants of those derivations such as

developers. In looking at a corpus, researchers instruct a computer program to search for all occurrences of a particular word based on the selected counting unit of either the lemma or the word family.

To illustrate a major criticism of word families in corpus-based research, consider the following example from the Corpus of Contemporary American English (Davies, 2008). Here are two occurrences of the word *developing* in context. First, *the ripple effects of international finance could turn nasty in a developing nation* and *teachers will be developing students' knowledge about medical technologies*. In counting word families, these two examples would both go under the frequency of *develop*. In counting lemmas, these sentences would count as one frequency for the adjective and one for the verb.

Historically, most vocabulary lists counted word families. The rationale is based on the assumed relative transparency of the connection between the inflected and derived forms. In other words, learners should be able to make connections when presented with the various forms in context such as reading, writing, and speaking (Bauer & Nation, 1993). Oft cited support for this assumption comes from Nagy, Anderson, Schommer, Scott, & Stallman (1989) who looked at the relationship between speed of word recognition and inflectional, derivational, and non-morphological related words in a study of U.S. university students ($n=95$). The researchers found a significantly faster recognition time for words that had an inflectional or derivational relationship such as *govern* and *government* but not between words that had a non-morphological relationship such as *feet* and *feel*.

More recent studies challenge this assumption and increasingly call for lemmatized vocabulary lists. Schmitt and Zimmerman (2002) conducted a study on the ability of non-native

English speaking (*n=106*) students at three levels of study to produce noun, verb, adjective, and adverb forms of 16 randomly selected headwords from the AWL. Results showed that the three groups only produced an average of 58.8% of the possible derived forms. Advanced ESL students at an IEP and ESL students taking an undergraduate writing course at a university could produce two to three of the derived forms. ESL students enrolled in a master of TESOL program could produce an average of three to four forms. Students were more likely to know noun and verb derivations as opposed to adjective and adverbs. The researchers also found a positive relationship between the ability to produce derived forms and a learners' overall reported vocabulary size. The researchers concluded that "teachers cannot assume that learners will absorb the derivative forms of a word family automatically from exposure" (p. 163). In the four word lists presented below, there is a shift from word family-based to lemmatized counting units.

General Vocabulary Lists

General Service List. The General Service List (GSL) (West, 1953) identified the 2,000 most useful word families in English from a corpus of 2.5 million words. The corpus included language samples from encyclopedias, textbooks, magazines, essays, novels, poetry, and science books. The researcher considered frequency, range, potential learning effort, necessity, register, and emotional connotation. By combining these factors, the GSL includes highly frequent words and those that are less ubiquitous but irreplaceable nonetheless. For example, West (1953, p. ix) explains the rationale behind including the word *preserve* on the GSL despite its relatively low frequency because it encompasses concepts such as *bottling, salting, freezing,* and *canning*, and it is not easily substituted by a higher frequency equivalent. The list is important for academically-minded learners for two reasons. First, Baumann's revision of the General Service

List (1995) ranked the word families by frequency creating two frequency bands. The first 1,000 words of the GSL, on average, account for 75 to 80% of running words in any text while the second 1,000 words cover between 4 and 6%. Secondly, the GSL is used to operationalize the level of difficulty of a language passage as in the development of graded readers.

New General Service List (NGSL). In 2013, Browne, Culligan, and Phillips revisited the concept of a general service vocabulary list by looking at a larger, modern corpus and retooling the definition of a general service word. The researchers used a subsection of the Cambridge English Corpus totaling 273 million words from nine different registers: learner language, fiction, journals, magazines, non-fiction, radio, spoken, documents, and TV. The researchers identified the most common words from the corpus to create the New General Service List (NGSL). The researchers identified 2,368 word families using a "modified lexeme approach" (Brown, 2013, p. 3). This approach counts a headword in all parts of speech and the inflected forms for each part of speech. The modified lexeme approach does not count derived forms from non-inflectional affixes. For example, occurrences of *list, lists, listed, listing,* and *listings* are counted but *un*listed is not (Brown, 2013, p. 3). The resulting list identifies almost 400 new word families which covered 90% of the NGSL corpus compared to 84% by the GSL. An increase in coverage is not particularly interesting as the additional word families in the NGSL would logically increase the coverage rates. What is noteworthy is that when the researchers lemmatized the NGSL, the list actually contains fewer words than the lemmatized GSL (2,818 vs. 3,623). In short, the researchers expanded the generalizability of the list while at the same time reducing the size of individual words.

New General Service List (New-GSL). Around the same time the NGSL was released, researchers from the United Kingdom published a New General Service List, hereafter termed New-GSL, which identified the most common lemmas in a corpus of over 12 billion running words (Brezina & Gablasova, 2013). The corpus consisted of subsections from the Lancaster-Oslo-Bergen Corpus (1 million words), the British National Corpus (100 million words), the BE06 Corpus of British English (1 million words), and the EnTenTen12 Corpus (12 billion words). The samples represented both written and spoken English in a variety of registers and disciplines. The final list includes 2,494 lemmas and provided coverage for an average of 80% of running words in the sample corpus. While the coverage rate is lower than the NGSL discussed in the previous paragraph, the New-GSL does provide empirical support of the existence of a general vocabulary. The researchers found that 70% of each of the four sub-corpora was accounted for by the core vocabulary in the New-GSL. In other words, the New-GSL items were equally represented across language samples of various sizes, modality, and discipline.

Academic Vocabulary

Academic Word List. The Academic Word List, or AWL, (Coxhead, 2000) continues to influence the methodologies of corpus-based word lists, materials for second language learners, and evaluation tools for academic writing. The 570 word families represented on the list are considered by many as the fundamental single-item words all academically-oriented ELs must acquire. The AWL is often used in tandem with the GSL to evaluate the vocabulary load of particular texts or to analyze student writing samples.

The AWL is based on a corpus of approximately 3.5 million words with a balanced proportion of art, business, law, and science. These four disciplines were further divided into 28

subfields, 7 in each discipline. The language samples also came from academic texts such as journals, university textbooks, and laboratory manuals. The scientific section of the corpus was supplemented by sub-sections of the Wellington Corpus of Written English, the Brown Corpus, and the Lancaster-Oslo Bergen Corpus. In all, 400 authors are represented in the AWL corpus in short (2,000 to 5,000 running words), medium (5,000 to 10,000 running words), and long (10,000 plus running words) text samples.

Coxhead excluded any word family from the GSL (West, 1953) for consideration on the AWL, which is one of the criticisms of the list and is discussed in the following section. The AWL word families occur in all four of the disciplines, are present in at least 15 of the 28 sub-fields, and have an overall frequency of 100. The combination of frequency and range criteria ensure that the results are generalizable to a wide range of fields. In a post hoc analysis, the AWL provided 10% coverage of running words in the academic text samples from the corpus used in the study, which support Nation's (2001a) 10% academic vocabulary estimate. In a second comparison, Coxhead analyzed a separate corpus of 678,000 words of academic language and found 8.5% coverage provided by the AWL. In a comparison corpus of non-academic language, the AWL provided less than 2% coverage. The discrepancies of coverage between the academic and non-academic samples provide support for the validity of the AWL in representing typical, general academic language in a wide range of disciplines.

Criticisms of the AWL. One component of Coxhead's operational definition of academic vocabulary automatically excluded any GSL word from inclusion on the AWL. The rationale being that general academic vocabulary is the next step up from general vocabulary; in other words, they are mutually exclusive bands of vocabulary. Comparing the coverage rates for the

first and second band of the GLS and the AWL should, consequently, show progressively lower percentages (Nation & Webb, 2010). However, the first band of the GSL covers 80% on average, and the second band provides between 4 and 6% coverage. The AWL provides around 10% coverage. Gardner and Davies (2013) also looked at the frequency of occurrence of the AWL in the Corpus of Contemporary American English and found that the words were redistributed in different frequency bands. Gardner and Davies suggested that the AWL is a subset of general, high frequency vocabulary and not a mutually exclusive set.

The AWL also became a lightning rod for critics against the very existence of a general academic vocabulary and in favor of a disciplinary approach. Hyland and Tse (2007; 2009) argue that there is no one-size-fits-all academic vocabulary, and the identification of important items should come from a close up view of the discourse patterns in respective disciplines. Only through disciplinary language analysis can a researcher see the salient words and associated meaning(s) relevant to any particular learning environment.

Hyland and Tse (2007) compiled a corpus of academic writing in science, engineering, and social sciences to evaluate the coverage rates provided by the AWL. The researchers found that the GSL and AWL combined provided 78% coverage of running words in the text. While Hyland and Tse use this figure as support for their hypothesis, the figure is only marginally lower than that reported in Coxhead's original study, 79.8%. A stronger criticism levied by Hyland and Tse, however, is the unequal distribution of the AWL word families across disciplines. According to their study, 534, or 94%, of the AWL word families have irregular frequency patterns across their science, engineering, and social studies sub-corpora. Of those 534 words, 227 had at least 60% of all occurrences in just one discipline. In other words, these AWL

word families would not meet the range requirements delineated in Coxhead's original study. Hyland and Tse concluded that the GSL and AWL word families would not give all learners proportional growth in language skills as they are used unequally across disciplines.

New Academic Vocabulary List. The New Academic Vocabulary List (AVL) (Gardner & Davies, 2013) is derived from 120 million words of written academic language from the Corpus of Contemporary American English. The researchers operationalized academic writing as that of academic journals, academic/trade magazines, and finance sections of the newspaper. The AVL counts lemmas instead of word families.

The researchers used four measurements to generate the AVL. First, the frequency of a word had to be 50% higher in the target corpus than a comparison corpus of non-academic writing. This measurement separated general high frequency and academic high frequency. Second, the researchers used a modified range criterion. Words must occur in seven of the nine disciplines represented in the corpus. To counter earlier criticisms regarding unequal occurrences in different disciplines, words also had to have at least 20% of their expected frequency in each discipline.

In calculating the expected frequency, the researchers provide an example with the word *ancestor*. The word *ancestor* occurs 3,596 times and covers .00297% of all running words in the academic sub-corpus. If *ancestor* were to be completely equally distributed in the education section of the academic sub-corpus, it would occur 238 times, or .00297% of 8,030,324 running words in education. The word only needed to meet 20% of that expectation to satisfy the frequency requirement. *Ancestor* only occurs 14 times, or 5.8% of the expected frequency and was excluded from the list.

The third criterion measures the dispersion of a word among the disciplines in the corpus to complement the range requirement. The researchers used the Juliand D figure (Juliand & Chang-Rodriquez, 1964) set at a .80 (possible scores range from .01 to 1.0) to measure how uniformly a word occurs throughout the entire corpus to avoid words that are very closely associated with any particular discipline and have a disproportionately high occurrence in that sub-corpora. Finally, the researchers created a disciplinary measure to further exclude technical vocabulary from the AVL. The researchers eliminated any word that occurred more than three times the expected frequency in any of the disciplinary sub-corpora.

After analyzing the corpus using these criteria, the researchers identified 2,000 academic words. In order to compare the AVL to the AWL, they retroactively assembled the list into word families and randomly selected 570 of them to compare coverage rates. The AVL covered 13.8% of the COCA academic corpus compared to 7.2% by the AWL. In the British National Corpus, the AVL covered 13.7% of the academic language section while the AWL covered around 6.9%. The AVL provides a strong defense against disciplinary-only critics.

Supporting the Existence of a General Academic Vocabulary

Coxhead generated the AWL over a decade ago, and it is still used in research, TESOL preparation programs, and English language materials. In building upon the research of the time, Coxhead identified the most common items in a sample of academic vocabulary while controlling for range of occurrence. Developments in technology and corpus linguistics have generated more sensitive tools to control for frequency and range requirements. The AVL, a modern interpretation of the AWL, does not discredit the AWL, but rather supports it by applying these more sensitive measurements (Gardner & Davies, 2013). The relationship

between academic and disciplinary vocabulary is that of increasingly specialized sub-sets, not separate entities (Gardner & Davies, 2013, Nation, 2001a). It is unwarranted to assume that support of general academic vocabulary discounts the existence of a discipline specific lexicon. Anecdotal support of this layering effect can be found in the very structure of higher education as students move from a generalist study of the liberal arts in their first two years of school to specialized courses within their respective fields of study. The discipline-specific vocabulary expectation continues to grow for students who choose to pursue a master or doctoral program. Vocabulary, whether operationalized as general, academic, or disciplinary is valuable in preparing academically-oriented language learners.

Formulaic Language

Krashen (1989) pointed out that travelers carry dictionaries instead of grammar books as a communication lifeline. This researcher, however, wishes to take the example one step further by pointing out that it is not only dictionaries but phrase books on which non-native speakers rely. Phrase books offer "islands of reliability" (Dechert, 1984, p. 227) that are situationally-appropriate pre-fabricated strings of language. For academic situations, these phrases are also important. Gibbons (1991, p. 3) contrasts the subtle vocabulary shifts in academic language: *If we increase the angle by five degrees, we could cut the circumference into equal parts* is somehow more correct than *If we make the angle five degrees bigger, we could cut the length of the outside of the circle into parts.* Vocabulary is important, but it is not important only on the word level but also the packages within which these words occur. From a theoretical standpoint, this is termed formulaic language and is the central construct in this study.

Vocabulary is not necessarily a group of letters surrounded by white space on either side (Folse, 2004; Halliday, 2004). Wray (2002) argued that "words do not go together, having first been apart, but, rather, belong together, and do not necessarily need separating" (p. 212). Pawley and Syder (1983) view the lexicon as a collection of units, not individual words, to be retrieved whole from long term memory. Yet there is no unifying definition and framework for the identification of a formulaic sequence. In fact, in a survey of literature, Wray (2002) found over 50 terms, many accompanied by unique operational definitions and identification techniques, to describe formulaic language. The most general conceptualization of formulaic language is:

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

The construct of a pattern-based language system has important implications in understanding second language acquisition.

Theoretical Framework of Second Language Acquisition and Formulaic Language

The current study is rooted in N. Ellis' connectionist theory (1998; 1999; 2003; 2005) of second language acquisition. In connectionism, language is seen as a system of complex, socially-embedded data. The data occurs in frequent patterns that learners store as chunks. As a learner is exposed to more and more language, these patterns are reinforced and their place in the developing linguistic system is solidified.

The purpose of storing patterns is twofold. First, the ultimate goal of language acquisition is to communicate. Connectionism views patterns as reliable avenues to communication.

Second, patterns provide learners with raw material to extract the underlying linguistic rules. Patterns also enable creative language production as acquired patterns are combined in new ways or filled in with new constructions. Creative language is generated through the interaction of implicit and explicit learning mechanisms.

In connectionism, both implicit and explicit learning play a role in SLA. Implicit learning is that which does not involve the learner's conscious attention or awareness. Explicit learning can include both inductive and deductive scenarios which draw attention to a rule and how it can be applied. In this model of SLA, implicit learning is connected to the frequency of input and frequency of learner output, or practice. Learners are sensitive to, but do not consciously count occurrences in either receptive or productive scenarios. However, with each exposure, implicit knowledge grows.

The role of implicit learning is not at the expense of explicit learning. Explicit instruction aids SLA by calling attention to previously unnoticed patterns. Once pointed out, learners can begin to collect frequency information through implicit learning. In this way, R. Ellis (2008) notes that connectionism is congruent with the weak version of Schmidt's (2001) noticing hypothesis. The noticing hypothesis states that SLA is facilitated when learners attend to particular features of input. Features that go unnoticed in input are stored in short term memory, but until the learner applies a direct focus to a particular linguistic item, its development will remain limited.

Just as implicit and explicit learning take on different roles in connectionism, so does the role of implicit and explicit knowledge. Implicit knowledge is intuitive and not able to be verbalized while explicit knowledge can be directly retrieved and reported. N. Ellis proposes a

weak interface between implicit and explicit knowledge in connectionism. He suggests that explicit knowledge helps generate creative language and can kick starting the implicit learning process which develops implicit knowledge. Explicit knowledge, however, cannot become implicit knowledge after time.

Building on the connectionist theory of SLA, Wray (2002) proposes a cognitive model of how these patterns are isolated, stored, and retrieved called the heteromorphic distributed lexicon (HDL). In the model, Wray conceptualizes an interaction between functional categories of vocabulary on three-tiers of lexical unit size. To illustrate the HDL, Wray provides five categories of vocabulary termed grammatical, referential, interactional, memorized, and reflexive. The functional categories are not of importance to this study nor in understanding the role of formulaic language in language production. The innovation in Wray's HDL model comes in the three levels of meaning analysis and how these levels interact.

In her HDL model, vocabulary is described as a three-tiered system of lexical units termed morphemes, formulaic words, and formulaic word strings. The first category, morphemes, includes a traditional account of the morphological system through both free (*over, happy, yes)* and bound (*-ly, -able)* varieties. Wray operationalized formulaic words (*established, maybe)* as polymorphemic words that are represented in the mental lexicon as a complete unit. The final category, or formulaic strings, are groups of words that are also processed holistically such as *in order to*, *half past [number 1-12],* and *the most important thing is [noun/verb phrase]* (Wray, 2002, p. 263).

The items that fill in these three levels are based on a learner's identification of the smallest unit of meaning that needs to be, as opposed to can be, segmented and analyzed (Wray,

2002). In other words, how learners extract meaning from input will determine the size of the unit stored in the lexicon. In some cases, the learner only needs to establish a link between a series of words and its semantics. Other times, sequences may be broken down into derivational and inflectional affixes. The concept of formulaic words is the most difficult to conceptualize. However, as mentioned earlier in the chapter, derived forms of words such as *happiness* often require explicit attention in order to produce accurately, and knowledge of one derivational form does not guarantee knowledge of another (Schmitt & Zimmerman, 2002). Furthermore, parts of speech tend to have unique collocation patterns that generate nuanced meanings (Stubbs, 1995). HDL extends these findings by suggesting the lexicon is composed of independent morphemes at three lengths based on the unit of meaning salient to a learner in their current stage of lexical development. Wray (2002) explains the multiple storage option as one that:

> Lists only those units…which direct experience has identified as communicatively useful…the effect is to prioritize the likelihood of the expected over the potential for the unexpected since the 'expected' is…that which the individual has observed to occur most of the time (p. 268).

By allowing for different levels of vocabulary as the base of the lexicon, Wray argues that the model more accurately represents the cognitive threshold for lexical irregularity. Wray (2002, p. 366) gives another example with the sequence *by and large*. This expression acts as an adverb as in *the flights at metropolitan airport <u>by and large</u> arrive on time.* It is not helpful to break down the sequence into three free morphemes *[by] + [and] + [large]* because it does not help decode the expression. Therefore, Wray argues that *by and large* is stored holistically as one unit in the lexicon at level three. Wray also allows for the individual free morphemes in the

expression to be stored in the first morphemic level. Repetition of storage is a key element in the HDL. The same three words (*by, and, large*) can be stored as level one morphemes and as a level three formulaic expression.

There are two potential weaknesses to the HDL model. First, the model does not have a traditional sense of efficiency (Wray, 2002). There is repetition in the mental lexicon by storing both the morphemes and the formulaic strings of language as separate units. However, Wray argues that this is actually a strength of the model as items are stored, processed, and recalled in the form which they most often appear and intend to be used. Also, as the lexicon is based on units of meaning needed by the learner, both rule and meaning-based utterances can be accounted for. Consider the following: *she always wears a gold ring on the other hand.* In this example, *on the other hand* is produced as a result of rule-based knowledge and has a different meaning from the formulaic-based expression of the same form that means *however.* By allowing the levels to interact, Wray argues that the HDL can account for both formulaic and non-formulaic language production. The model also connects with the work of Hoey (2005) who noted that the study of vocabulary through corpus linguistics is only able to provide a potential representation of the input an individual learner receives. Large corpora will never be able to fully represent a person's cognitive corpus; therefore, allowing formulaic language to be determined by a learner's need is important.

The biggest weakness with the model is the grey area between the morpheme and formulaic word categories. For example, Wray lists words like *maybe* and *because* as formulaic words (p. 263), but these do not seem as convincing as true compound words like *cellphone* or derivations such as *unbelievable.* While historically words like *maybe* and *because* were in fact

two free morphemes that gradually lost the separating space overtime, Wray acknowledges that their classification as formulaic words would be of more interest to a lexicographer than a second language learner . However, allowing for a fluid, non-discrete classification system helps theoretically connect vocabulary and grammar and eliminate the need to "place formulaic word strings awkwardly at the end of lexical models and, often, entirely outside of grammatical ones" (p. 261).

Methodological Framework for Studying Formulaic Language

There are two methodological camps in the study of formulaic language. The first is a phraseological approach which focuses on fixed lexical combinations and idiomatic expressions. Phraseologists use human judgment to qualitatively and deductively identify items (Granger & Paquot, 2008; Schmitt 2010) based on a spectrum of transparency and substitutability. An example of transparency and substitutability would come in the expressions *to reach a conclusion* and *to arrive at the findings* (Howarth, 1998, p.162). These verbs seem to have a clear, transparent meaning, but they are not interchangeable in the expressions. Each expression, therefore, is one phraseological unit. Phraseological studies typically produce output of two-word collocations, but longer expressions can be identified (Schmitt, 2010). One major obstacle to the identification of formulaic language in this approach is that results are difficult to replicate as the main criteria for identification is human judgment of degrees of semantic transparency and fixedness (Schmitt, 2010). Phraseological studies are based on intuition in the design and interpretation of results (Wray, 2002), which Sinclair (1991, p. 4) argues "…is highly specific, and not at all a good guide to what actually happens when the same people actually use the language."

The psycholinguistic approach emphasizes inductive corpus analyses by computer programs (Granger & Paquot, 2008; Schmitt, 2010). Psycholinguistic research makes extensive use of corpora to gather hard data on the occurrences of formulaic sequences (Altenberg, 1998; Moon, 1998) and to make comparisons between language used in particular situations of interest (Granger, 1998; Howarth, 1998). This approach also comes with criticisms. Frequency counts can be over-powerful by identifying expressions that have little real value or meaning (Wray, 2002). Human intuition is then needed to make judgments about the computer-based results; thereby generating the same weaknesses as the phraseological approach (Nattinger & DeCarrico, 1992). It can also be greatly skewed by 'fifteen minutes of fame' language trends (Wray, 2002). For these reasons, careful selection of language samples to build the corpus is critical, regardless if one is looking for single or multi-word vocabulary items.

The line between the two approaches is not stark. It is blurred as researchers from the phraseological tradition such as Moon (1998) and Howarth (1998) have utilized frequency counts as well as traditional psycholinguistic researchers such as Simpson-Vlach and Ellis (2010) who incorporated teacher perceptions in the Academic Formulas List. Regardless of the approach, both overlap in the view that ready-made combinations of language are ever present in both speaking and writing, and they have implications for the acquisition and processing of language by moving away from the slot-based, assembly line conceptualizations of language (Cowie, 1998; Harwood, 2002; Schmitt 2010; Schmitt & Carter 2004; Sinclair, 1991; Simpson-Vlach & Ellis, 2010; Wray, 2002). Both camps also advocate for an expanded understanding of how learners use formulaic language when compared to their native-speaking counterparts, and they reject the criticism that formulaic language "reduces language learning to phrase-book

memorization and repetition" (Howarth, 1998 p. 165). Each group views creative language and formulaic language as complementary constructs where formulaic language is the scaffolding from which creative language is built (Harwood, 2002; Howarth, 1998; Wray, 2002)

Examples of Formulaic Language

As evidenced in the heteromorphic distributed lexicon model (Wray, 2002), formulaic language comes in many different forms. The current study focuses on level three formulaic sequences, but a review of several varieties is presented here. Some of the shortest examples of the level three lexicon are collocations. Biber, Conrad, and Reppen (1998), in an analysis of 2.7 million academic words, showed that the combination *large number* occurred 48 times per million words as opposed to *great number* which only occurred around 9 times per million words. If looking at the two collocations in isolation, intuition would point to these groupings as equivalent, yet based on the figures, the former is somehow a more correct. Stubbs (1995) analyzed 3.5 million words of spoken and written English and identified common collocations such as *provide work* and *cause work*; however, he points out that these terms imply different attitudes, or semantic prosody. In *the new policy will provide work for thousands* and *the new policy will cause work for thousands*, the first sentence seems positive and supportive while the latter implies criticism. Schmitt and Carter (2004, p. 8) offer another example with the collocation *bordering on*. In an analysis of the British National Corpus, 27 out of 100 occurrences referenced a location such as *a property bordering on a busy road must have a sidewalk for pedestrians*. The most common use, by far, was actually quite different as in *the actions of the managers appeared to me as bordering on negligence*. In this second example, *bordering on* actually means 'approaching an undesirable state' (Schmitt & Carter, 2004, p. 8).

Sentence-like units are also formulaic phrases, or lexical bundles. One type of phrase is an open sentence frame where contextually-based information is added to make it relevant such as ___ *thinks nothing of* ___ (Schmitt & Carter, 2004, p.7). The purpose of the phrase is to express something unexpected as in *my roommate thinks nothing of staying up until 3am working*. Schmitt and Carter, however, note that while there are two open slots in the phrase, there are not limitless options for filling them. The sentence *my roommate thinks nothing of going grocery shopping* sounds odd because the act of grocery shopping is not unusual or unexpected. Other examples of sequences that have a clear pragmatic function are *I'm sorry to hear about* __. This bundle is an acceptable way to express sympathy for bad news, while *I'd be happy to* __ can be used to accept a request for assistance (Nattinger & DeCarrico, 1992, p. 62-63).

Bundles are not necessarily incomplete. Whole sentences can be formulaic as in *hot today isn't it* to initiate small talk with a stranger. Even the question *how are you* and the response *I'm fine, thank you* could be deemed formulaic, as many non-native speakers are initially confused by the question. The pragmatics of the expression are equivalent to a greeting, not a genuine show of interest. Formulaic language can also be found in technical scenarios, even though it may not appear as formulaic to individuals outside a particular speech community (Schmitt & Carter, 2010; Wray, 2002). A new professor at a university talks about his or her teaching schedule as *a two-two load* and restaurant workers know not to sell any more surf and turf after they hear *eighty-six the lobster.*

Another possible form of formulaic sequences are those used for discourse organization. Examples of bundles with a discursive function are: *the form of, the nature of, the role of, the*

structure of, the size of the, and *as a result of* (Biber et al., 1999; Biber et al., 2004; Simpson-Vlach & Ellis, 2010). Schmitt and Carter (2004) note that some of these bundles (*on the other hand, in conclusion*) are well known to ESL teachers, especially those that teach writing. Wray (2002) explains that here "language is being used to externalize relationships between ideas which are already in the mind of the speaker" (p. 87). These sequences are used to connect strings of language together in a cohesive manner. They can also provide information on the function of subsequent language, e.g. to support, to oppose, to conclude, etc.

In each of these examples, meaning is conveyed not through one word, but a series of words working together (Sinclair, 1991). In the discussion of language proficiency presented at the beginning of this chapter, Bachman (1990) noted that communicative competence involved drawing on grammar and pragmatic skills. Sinclair (1991) and Halliday (2004) propose that these systems are not mutually exclusive. Language does not have a grammar network that creates slots for content and function words to fill and generate meaning. Language consists of a lexiogrammar where syntax and vocabulary cannot be separated. When the subsystems of grammar and vocabulary intertwine, typical collocation patterns emerge. Language that incorporates these patterns is seen as more cohesive (Halliday & Hasan, 1976). In this way, the connectionist theory (Ellis, 1998; 1999; 2003; 2005) is congruent with the lexiogrammar perspective (Halliday, 2004; Sinclair, 1991) as language patterns are used to inductively decode grammar rules. In sum, from the examples in the previous paragraphs, it is impossible to determine if the formulaic patterns occur because words fall into certain sequences or certain sequences simply contain particular words (Bennett, 2010). What is undisputed in the research is the prevalence of these patterns in language production.

The Prevalence of Formulaic Sequences in Spoken and Written Language

Research has established the presence of formulaic language in both speaking and writing, but the actual concentration of occurrences varies. This is largely due to the variety of identification techniques used, as described previously. Watershed reports from both perspectives are presented.

Biber et al.'s (1999) lexical bundle project is one of the most influential studies on formulaic language to date. The researchers used the Longman Spoken and Written English Corpus (LSWE) containing over 40 million words divided into conversation, fiction, news, and academic registers. The number of texts included in the corpus is 37, 244. The academic prose subsection of the corpus includes 5,331,800 running words from 408 texts. The academic writing samples include 75 book extracts and 333 research articles. There are thirteen disciplines represented in academic writing.

In this study, the authors used lexical bundles to refer to formulaic sequences of three or more words that occur within a register a minimum of ten times per million words. These occurrences must come from at least five different texts within the register. As the researchers also allowed for larger strings up to five or six words, these longer sequences had a reduced frequency requirement of five times per million words.

The researchers found 1,500 bundles that met the frequency and range requirements listed above. The most common 3 word lexical bundles in academic writing, with over 200 occurrences per million words are: *in order to, one of the, part of the, the number of, the presence of, the use of, the fact that, there is a, there is no*. The most common four word bundles,

occurring over 100 times per million words are: *in the case of, on the other hand* (Biber et al., 1999, p. 994).

In the sub-corpus of conversational English, which contained 3.9 million words, the researchers found that 30% of conversation could be attributed to recurrent combinations of words. In an academic writing sub-corpus of over 5 million words, 21% of the samples were made up of recurrent strings of three or more words.

Biber et al. (2004) looked at four-word sequences in 1.2 million words of university classroom discourse from business, education, engineering, humanities, natural science, and social science undergraduate and graduate courses at four universities. They compared this corpus to the conversational sub-corpus used in the 1999 study described in the previous paragraphs. They found that classroom teaching contained a common core of 84 formulaic sequences, which was nearly double the 43 common sequences found in conversation. Biber et al. also looked at textbooks as representative of academic writing in a corpus of over 700,000 words and compared the sequences to those found in the academic writing sub-corpus from the 1999 study. They found a higher number of formulaic sequences in textbook writing, but the frequency of occurrence was higher in academic writing (3,500 per million words) than in textbook writing (2,000 per million words).

Altenburg (1998) analyzed the 500,000 word London-Lund Corpus of Spoken English. Altenburg found 68,000 recurrent two-word combinations covering 80% of the corpus. Altenburg, however, cautioned that many of these expressions identified by frequency alone produced nonsense strings of language such as *and the* and *in a.* When refined to three-word

combinations that occurred at least ten times, the number dropped to 6,692. While still a notable figure, it illustrates the caution is needed when interpreting results.

Howarth's (1998) study of academic written language in the social sciences found that formulaic language accounted for between 30 and 40% of two smaller corpora totaling around 200,000 words. Moon (1998) looked at 67,000 formulaic sequences in an 18 million word corpus of British English. The corpus was mainly assembled from journalism samples, which combines features of writing and speaking. Other language samples included non-fiction, fiction, and conversation. The formulaic sequences came from the Collins Cobuild English Language Dictionary (1987). Moon's results overall support previously presented studies in that formulaic language is indeed prevalent. However, instead of reporting the percent coverage provided by the sequences, she broke them down into frequency bands. Moon found that 32% of the sequences from Cobuild only occurred once in her corpus. This is important as five to ten occurrences per million words is an important threshold in corpus-based research (Biber et al., 1999; Biber et al., 2004; Simpson-Vlach & Ellis, 2010). Moon found only 4%, or 2,680, met this frequency threshold. In addition, less than 1% occurred between 50 to 100 times in the corpus.

These studies document the fluctuation in occurrence of formulaic language in both spoken and written contexts. Conklin and Schmitt (2012) suggest a reasonable estimate of formulaic language in discourse is between 30 and 50%. In response to these figures, researchers have attempted to identify and quantify the gap between native and non-native speakers' use of sequences, especially in writing. Overall, the results do show a significant gap in the productive use of formulaic language by language learners coupled with an over-reliance on particular sub-sets of sequences. Researchers speculate that these two factors contribute to

the ‘non-nativeness’ of second language writing. Before discussing the empirical research on the gap between non-native speakers’ use of formulaic language, it is important to dig a little deeper into why this gap matters. Evidence suggests formulaic language aids in fluent processing of language.

The Processing Benefits of Formulaic Language

Language input is initially stored in short term memory where, under a connectionist view of SLA (N. Ellis, 1998; 1999; 2003; 2005), frequency of input and practice are used to determine what information is important enough to transfer to long term memory. Estimates suggest that short-term memory can hold between eight to ten words and that formulaic language is prevalent because it acts as a booster to short term memory (Bower, 1969; Miller, 1956; Pawley & Syder, 1983; Simon 1974).

Arnon and Snider (2010) looked at the effect of frequency on the comprehension of four-word sequences that had a transparent meaning. They found processing times decreased as a function of frequency. Even though the sequences were not idiomatic, the words still functioned together as a group. Schmitt, Grandage, and Adolphs (2004) included corpus-identified formulaic sequences in a dictation activity consisting of 24 words for native and non-native speakers. When learners had to recreate the dictation, many, but not all, of the formulaic sequences were produced fluently, but there did not appear to be a clear pattern to indicate which sequences would lead to fluent reproduction. Schmitt, Grandage, and Adolphs conclude that the variance operates at an individual level depending on what is formulaic for the individual learner; which is congruent with the HDL (Wray, 2002).

Other studies have shown that formulaic language also aids in the production of fluent language. Kuiper (1996) studied language production of sport commentators and auctioneers as both scenarios involved time constraints and performance pressure. Kuiper found that sports commentators at horse races and announcers at cattle auctions used high amounts of formulaic language to structure their commentary. Pawley (1991) also studied formulaic language at horse races and cricket matches and found a very high concentration of formulaic language in play-by-play commentary. Crystal (1995) looked at formulaic language in weather forecasts and found a strong reliance on formulaic language.

Formulaic language allows both creators and recipients of language to focus on the discourse as a whole and not on the individual words (Kuiper, 1996; Oppenheim, 2000). Overall, there seems to be a connection between how routine a communicative scenario is, the amount of performance pressure, and the amount of formulaic language. The more often an individual communicates in a particular scenario or the less time they have to generate language, the less they will rely on generating language from scratch (Becker, 1975).

The Gap in Formulaic Language Production and Attempts to Bridge It

The defining factor for inclusion in this section was an analysis of formulaic language used by instructed adult second language learners. This section, therefore, draws upon studies of formulaic language from both phraseological and psycholinguistic perspectives. A dual perspective approach also requires the inclusion of studies that operationalize formulaic language at different lengths. As research related to pedagogical attempts at formulaic language development are sparse, studies are included that look at both native and non-native speakers.

Evidence of Formulaic Language in the Writing of Adult Language Learners

A recent study by O'Donnell, Römer, & Ellis (2013) offers reassurance that while different operational and identification approaches produce different results, the conclusions share some common ground. The research trio studied the development of formulaic language in writing between native and non-native speakers and expert and novice level writers as evidenced by four identification approaches: frequency-based, association-based, open frame-based, and native norm-based. Frequency and association-based measures both fit into a psycholinguistic understanding of formulaic language. Open-frames are sequences with a variable slot included. They are closely associated with Sinclair's (1991) idiom principle showing that not all words are equally possible when language is put in context. It is largely a psycholinguistic measure, but the semantic cohesiveness needed in open-frame identification moves it closer to a middle ground on the psycholinguistic-phraseological spectrum. The same shift can be said for native-norm approaches as native speaker intuition is needed for identification in addition to psycholinguistic measures of frequency and association.

O'Donnell, Römer, and Ellis created four test sub-corpora representing native speaking undergraduate novice writers, non-native speaking undergraduate novice writers, graduate student published academic writers, and seasoned published academic writers. Seasoned published academic writers were the only individuals missing information on the writer's L1. The researchers identified formulaic language according to four identification approaches and used a factorial ANOVA to compare the results (Table 2).

Table 1.

Differences in Formulaic Language among Writing Groups

	Experts vs. Graduate	Graduate vs. Undergraduate	Native vs. Non-Native
Frequency-based	=	↑	↑
Association-based	=	↑	--
Sentence frames	--	--	--
Native-norm	↑	=	--

Note. Summarized from O'Donnell et al. (2013)

The frequency-based analysis identified three, four, and five-word formulaic sequences that occurred a minimum of three times in the four corpora. The researchers found a significant difference based on writing group, $F(4, 345) = 73.46, p < .001$, sequence length, $F(2, 345) = 2781.40, p < .001$, and interaction between group and length, $F(8, 345) = 4.09, p < .001$. Using a post hoc Tukey analysis with a significance threshold of $p < .05$, the researchers determined that there was no difference between expert and graduate writers regardless of the graduate writers' status as a native or non-native speaker. For undergraduate students, there was a significant difference between native and non-native speakers in the MICUSP samples, but not between the LOCNESS and ICLE samples.

The association approach used a mutual information (MI) (Oakes, 1998) score threshold of 6.72 for three, 13.09 for four, and 20.84 for five-word sequences. MI is based on research in cognitive science but has been applied in the study of formulaic language (Schmitt, 2010; Simpson-Vlach & Ellis, 2010). Mutual information (MI) is a scalar variable that measures the strength of association between two independent words. It is not a test of significance. Mathematically, it is determined by the probability of words occurring together divided by the

probability of the words occurring in isolation (Oakes, 1998). These minimums are the mean MI scores for bundles of each length. The researchers found a significant difference in bundle use between writing group, $F(14, 315) = 8.59, p < .001$, sequence length, $F(22, 315) = 244.20, p < .001$, and interaction effect between group and length, $F(22, 315) = 4.42, p < .001$. The post hoc analysis showed no significant difference between expert writers and both native and non-native speaking graduate university writers in the MICUSP corpus, but these two groups used more formulaic sequences in their writing than samples from the ICLE and LOCNESS. There was however, no difference in bundle use between ICLE and LOCNESS samples.

When operationalizing formulaic sequences as sentence frames, the researchers found no difference in sentence frame use between writing group, $F(4, 345) = 0.94, p > .05$. Significant differences did occur for frame length, $F(2, 345) = 669.72, p < .001$, and interaction between group and length, $F(8, 345) = 35.53, p < .001$. A post hoc analysis, however, reconfirmed no significant differences between any groups of writers.

Native-speaker norm bundles came from the Academic Formulas List (Simpson-Vlach & Ellis, 2010) that grounds psycholinguistic measures in teachers' perceptions of bundle worth. This approach demonstrated significant effects for writing group, $F(4, 345) = 6.03, p < .001$. There was no measurement for length because the native-norm standard identified unequal groups of three, four, and five-word sequences. The post hoc analysis only showed one significant difference between seasoned academic writers and all other groups. There was no significant difference between native and non-native speakers.

The results of these four analyses indicate two important conclusions. First, the identification process affects the total number of sequences identified which influences the size

of the gap between native and non-native speakers and novice and expert writers. That being said, there does appear to be a trend between the frequency, association, and native-norm based analyses. Experience, for both native and non-native speakers, has the greatest effect on the use of formulaic language. A look at the earlier work of Yorio supports these conclusions.

Yorio (1989) conducted multiple analyses on formulaic language in adult second language writers. The first, a longitudinal case study, analyzed 14 writing assignments from student K completed as part of a university course. K was an 18 year old native Korean speaker living in the New York metropolitan area. He attended middle and high school in the United States, but only enrolled in one ESL course during his first semester in middle school. His secondary educational background was largely in mainstream classrooms. Yorio's initial focus of the study was an analysis of K's syntax and morphology development. He used t-units (Hunt, 1964) as the unit of analysis which consists of one main clause and associated subordinate clauses. In the morphosyntactic analysis, Yorio found that only 32% of the t-units were error free; but 98% were "perfectly comprehensible" (p. 60). Upon further investigation, he found that K utilized a large number of idioms, set phrases and patterns, and collocations in his writing. Yorio concluded that the use of these lexical conventions were responsible for the comprehensibility despite the morphological and syntactical errors in writing.

In Yorio's second study, he compared the writing of non-native (*n=25*) and native speakers (*n=15*) studying at the same college in the United States who failed a university-wide requirement of writing proficiency. The non-native speakers had lived in the United States between five and seven years. Yorio conducted a lexical and error analysis on the students' writing assessments. Overall, native speakers and non-native speakers differed in their use of

collocations and formulaic expressions. Native speakers used 145 collocations that made up 19.5% of their writing. Non-native speakers used 138 collocations representing 14% of their writing. From these 138 collocations, a little over half were produced without any error in word form or grammar. Formulaic sequences showed the most disparity. Native speakers used 52 formulaic expressions, or 36% of their writing, while non-native speakers only used 9. This represented 6.5% of non-native speaking writing, but non-native speakers showed much more form and grammar control compared to collocations. Yorio found that there was a relationship between non-native speaker proficiency level and collocation and formulaic sequence use. The higher the proficiency level, the more formulaic language was used. He describes this relationship as "non-trivial" (p. 64) but does not give any statistical support. Yorio concluded that lexical patterns greatly contribute to the idiomaticity, or the "native like quality which….is a non-phonological 'accent', not always attributable to surface errors, but to certain undefined quality" (p. 64).

In Yorio's final study, he looked at two groups of non-native speakers' performance on the same writing assignment. The first group consisted of learners in an ESL class in the United States. The second group were English majors at a university in Argentina. The ESL learners had been living in the United States for five to six years. All were native Spanish speakers. The EFL learners/English majors had studied in their home country for three to five years, and none had visited an English-speaking community. He hypothesized that the EFL students would outperform ESL students in terms of grammatical accuracy, but the ESL students would be superior in terms of idiomatic language. In terms of grammar, his hypothesis was correct. EFL students showed fewer grammatical errors in their writing. However, they also surpassed the

ESL students in terms of idiomatic, formulaic language. Yorio concluded proficiency overall is the most important indicator of formulaic language use. He draws upon Pawley and Syder's (1983) distinction between native-like selection and native-like fluency to support the conclusion. Native-like fluency is the ability to produce seamlessly connected strings of language while selection relates to using natural-sounding constructions. Other studies take a closer look at exactly what kind of differences are included in the formulaic language gap.

Howarth (1998) investigated the occurrence of non-standard phraseology in ten essays written by non-native speakers at the end of their first semester in an MATESOL program. The researcher identified target formulaic sequences consisting of verb and noun complements that occurred ten or more times in the Lancaster-Oslo-Bergen and Leeds University corpora. The analysis resulted in 63 lexical units classified by the researcher as either free or restricted (Howarth, 1998, p. 28). Items in a free unit retain their literal meaning as in *blow a trumpet*. In a restricted unit, the word that is matched with the verb has a specialized lexical function as in *blow a fuse*. Howarth found that native speakers use about 50% more restricted units and idioms than non-native speakers in writing.

Another study by Hewings and Hewings (2002) investigated the use of it-clauses in 15 dissertations written by non-native speakers in the field of business. The researchers also used a small comparison sample of 28 academic journal articles as a model of expert level writing, regardless of the authors' linguistic background. The researchers operationalized formulaic language as any it-clause where the subject of the sentence was placed at the end of the clause. The researchers found that student writers used 38% more it-clauses than the comparison sample of journal articles. Upon further investigation, Hewings and Hewings reported the primary

functions of overused it-clauses were attitude, emphasis, and attribution of characteristics. The sentence *it should be noted that the findings did not support the hypothesis* would be an example of an emphasis clause. Non-native speakers, however, showed 17% less hedging in their writing as demonstrated by it-clauses. The researchers suggest that non-native speakers use more tools of persuasion in their writing, regardless of the factual basis of the argument. The researchers note the limitation of using dissertations as language samples for analysis because the final draft would have received feedback from multiple faculty members, who may or may not be native speakers, and involve multiple revisions. They argue that since the purpose of the study is not error analysis for the purpose of understanding the acquisition and errors of vocabulary, it is appropriate to analyze stylistic differences that persist.

Granger (1998) looked closely at specific structures in second language writing. She investigated formulaic sequences in the writing of advanced English learners who spoke French as a first language. Granger outlines the target sequences as (p. 8):

(1) Passive sentence frames:

It + (modal) + passive form of a saying or thinking verb + that-clause

Ex: *It is thought that…* OR *It could be said that…*

(2) Active sentence frames:

Pronoun + (modal) + active form of a saying or thinking verb + that-clause

Ex: *I maintain that…* OR *We could say that…*

The researcher found that native speakers use the passive sentence frame 77 times per 200,000 words compared to non-native speakers at 52 times. The active sentence frame was used in native speaker writing 109 times per 200,000 words. Non-native speakers used the

active frame 399 times per 200,000 words. The most overused sequences in non-native speaker writing were (p. 9):

(1) Pronoun + can/cannot/may/could/might + say + that (75 NNS vs 4 NS)

(2) I think that (72 NNS vs 3 NS)

(3) Pronoun + can/could/should/may/must + notice+ that (16 NNS vs 0 NS)

(4) Pronoun + may/should/must + not forget + that (13 NNS vs 0 NS)

While previous research argued that a lack of formulaic language makes writing seem non-native like, Granger's study shows that an overuse of sequences can contribute to a perception of wordiness or lack of clarity of ideas. She cautions that learners might "cling on to certain fixed expression which they feel confident in using" (p. 10) which suggests that sequences can at some point become less of a life raft and more a part of the storm.

Finally, Durrant and Schmitt (2009) studied the smallest measure of formulaic language, two-word collocations. They looked at frequency-based adjective/noun and noun/noun word pairs in non-native and native writing samples. Non-native writers in this study were graduate students in a pre-semester English for Academic Purposes (EAP) course and undergraduate students enrolled in an EAP course during their first year of study. Native writers were graduate students in an applied linguistics program and essays taken from a current affairs magazine. The researchers collected short (*n=24*) and long (*n=24*) text samples from native writers as well as short *(n=24*) and long (*n=24*) samples from non-native writers. The researchers used the British National Corpus World Edition to identify target word pairs based on frequency and mutual information score (Oakes, 1998).

Results indicated two important differences between native and non-native writers' word pairings. First, on average non-native writers used significantly fewer low frequency word combinations (*M=38.14, SE=3.39*) than native writers (*M=48.19, SE=1.52*), *t(46)=3.55, p<.001, r=.46*. Second, non-native writers tended to use fewer word combinations with a mutual information score higher than 8 (*M= 14.95, SE= 1.46*) than native writers (*M=17.48, SE=1.30*) in writing (*t(46)=3.39, p<.001, r=.45*). The researchers conclude that non-native speakers are conservative in their use of low frequency word combinations and prefer combinations that are frequently represented in input. The researchers also suggest that these findings support a second language learners' ability to acquire phraseology in the L2, but that they are more likely to acquire frequent combinations as opposed to other salient forms of word pairings.

The studies described in this section all point to a performance gap between native and non-native speakers' use of formulaic language in writing. The research also hints that the gap might not solely be the result of developing language proficiency. Experience as a writer was an important precursor to formulaic language use. While there was certainly a difference between native and non-native writers' quantity of sequences, analyses of apprentice and experienced writers also showed a disparity. In other cases, it was the quality not the quantity of formulaic language that marked non-native writing. Second language learners, even at the advanced level, tended to overuse particular sequences in their writing. It remains unclear how this gap in performance affects second language writing scores. Only one study has directly investigated this topic.

Ohlrogge (2009) looked at the relationship between formulaic language and L2 writing score for intermediate-level non-native speakers (n=170) on an in-house proficiency test used at

the University of Michigan's IEP. Ohlrogge operationalized formulaic language based on Wray's (2002) HDL model. Using the morpheme-equivalent approach, he generated eight groups of sequences: collocations, idioms, phrasal verbs, personal stance markers, transitions, language borrowed from the writing prompt, and generic rhetorical patterns and biographical information. Five of these categories warrant further explanation.

Personal stance markers included strings such as *from my perspective* that occur at the beginning of a sentence. By transitions, Ohlrogge refers to discourse organizing sequences such as *first of all* and *on the other hand* that connect the text. Any string of words directly copied from the prompt in the writing assignment were also categorized as formulaic. These sequences consisted of multiple phrases with minimal morphological changes from their original form. Generic rhetoric patterns include those that "the writer has had time to perfect in advance of the actual writing assignment...[they are] standard openers and closers" (p. 380). Ohlrogge gives an example of these longer sequences with *taking all of the above into consideration, it should come as no surprise that* (p. 380). Finally, the researcher deemed any biographic data used to establish the writers identity in the essays as formulaic because they are routine productions for that individual speaker.

The researcher and two additional readers identified sequences independently. Once coded for formulaic language, the essays were grouped into five mini corpora based on score from a five-point holistic rubric. All scores were determined independent from Ohlrogge's study during the original scoring session by trained raters. The researcher first calculated the frequency of each formulaic category by score. Ohlrogge found that idioms, collocations, transitions, and personal stance markers did increase in frequency between the lowest and highest score level,

but not at a constant rate. Copied language from the prompt, biographic data, phrasal verbs, and generic rhetoric all decreased between lowest and highest-rated essays, but also at different rates.

Using a Spearman rank correlation, Ohlrogge calculated the relationship between each type of formulaic language and essay score. Only three types of sequences significantly correlated ($p<.05$) with score: idioms and collocations (.90), personal stance markers (.90), and copied text from the prompt (-.82). Ohlrogge found inconclusive results relating to the correlation between transitions (.70) and score. Phrasal verbs (.60), generic rhetoric (-.10), and biographic information (.10) did not significantly correlate with score.

In conclusion, Ohlrogge found that different proficiency levels used different types of formulaic language in different ways, but that formulaic language use generally increased with proficiency level. However, only idioms, collocations, and personal stance markers positively contributed to essay score. Ohlrogge concludes that direct instruction of formulaic language would be beneficial for L2 learners. He also concludes that more research is needed to understand to what extent and in what way raters are influenced by the use of formulaic language in writing. However, one criticism of Ohlrogge's study is the operationalization of formulaic language. While relying on the HDL theoretical model (Wray, 2002), the identification approach using only human judgment makes replication of the identification results difficult.

Ohlrogge's study also supports an earlier study by Hawkey and Barker (2004). While the goal of this particular study was not connected with formulaic language per se, the researchers were trying to identify distinguishing features of non-native writing on Cambridge ESOL examinations. In an examination of a small set of 98 writing samples totaling 18,000 words, the researchers did indicate that collocations and idioms occurred more frequently in higher rated

essay bands, but that repetition of certain phrases was a hallmark of essays in lower proficiency bands.

Evidence of Formulaic Language in the Speech of Adult Language Learners

While the current study investigates formulaic language in academic writing, there are three studies that warrant discussion on the perceived benefits of vocabulary strings to improve fluency for second language learners. Wood (2006) studied the effect of formulaic language on the fluency of ESL students (*n=11*) at a Canadian university. Wood operationalized improved fluency based on two measurements. First, he looked at the total number of syllables in a recorded speech sample divided by the total number of pauses. Second, he determined the total number of formulaic sequences divided by the total number of segments, or sections of speech between two pauses. In other words, Wood tested if using formulaic language increased the length of fluent speech segments, which he described as improved fluency. Data was collected over a six month period. Each learner watched three short silent films twice with a gap of two months between each viewing. After viewing the film, the students were asked to retell the plot without the aid of notes or outside resources.

Three native speaking graduate students in applied linguistics used their judgment to identify formulaic language in each speech sample. For the purposes of Wood's study, a sequence was considered formulaic if it was identified by at least two of the three judges. Judges used five criteria when reviewing the speech samples, and they could identify sequences that contained one, some, or all of the following characteristics. First, a sequence could be phonologically different from the surrounding discourse by using a continuous and coherent intonation pattern without pauses. The articulation may be faster than the surrounding discourse,

and at times even sound muddled as learners overlook individual words within a sequence. Second, judges looked for a group of words that were markedly different in length and complexity than the surrounding speech. Wood gives an example of a learner who used the expression *I would like to* but never used the modal *would* outside of this expression (p. 22). The next two criteria connect to the meaning and structure of a potential sequence. Judges pinpointed sequences that were idiomatic and/or had a restrictive syntactical structure. For example, the expression *beat around the bush* does not have a transparent meaning, and it cannot take a variety of grammatical forms, e.g. the expression *beat around the bushes* is not permitted in English. Finally, any formulaic sequence included in Nattinger and DeCarrico's (1992) study was marked as formulaic.

The results indicated that, as a group, the 11 participants increased their fluency and use of formulaic language between the first and second retelling of each film. In two months, the mean length of speech segments increased from 3.6 syllables to 4.2 for film 1, 3.6 to 3.8 for film 2, and 4.1 to 4.3 for film 3. The average number of formulaic sequences used also increased in each speech segment for film 1 (.27 to .40), film 2 (.31 to .37), and film 3 (.39 to .41).

Wood also analyzed the transcripts to categorize how formulaic language, as identified by the NS judges, was used to increase the length of speech segments. He identified five ways formulaic language improved fluency. One typical use involved simply repeating a formulaic sequence to extend the length of an utterance before a pause. Wood gives the following example: *The dynamite ah (pause) ah caused a (pause) motion where the man died (pause) and when he's <u>really happy</u> he's hysteric he's <u>really happy</u> because (pause) he's <u>really happy</u>* (p. 25). In this sentence, the longest speech segment contains the same two-word collocation repeated twice. A

second category included the stringing together of multiple sequences to extend the length of speech. In the sentence *He's make music by himself in his room* the NNS used three sequences identified by the judges in quick succession (p. 26). Wood termed the third use of formulaic language as "an illusion of increased fluency by relying heaving on one simple formula" (p. 26). One NNS in the study used the sequence *in the middle* repeated over and over in lieu of taking a pause. *In the middle of (pause) the land (pause) in the middle or between in the middle of (pause) the land (pause) in the middle of between continues in the middle (pause) the middle of their houses* (p. 27). The fourth use of formulaic language is similar to the third, except the repeated sequence is one identified with "self-talk" such as *I don't know, I think that,* or *I guess.* The final category of formulaic language use involves cases of organizational sequences to help build and move the narrative from point to point. Judge-identified sequences in this category include *beginning of the, the start of the history, go ahead,* and *this is the end of the story.*

In the end, Wood concluded that the increase in formulaic sequences between the first and second retelling of each story augmented the fluency of the NNSs because the length of the utterance between pauses was extended. However, in looking at the examples Wood provided for each category, the formulaic sequences do not seem to be adding cohesiveness of meaning. Wood's judges were only tasked with identifying sequences, and they were not asked to evaluate the actual quality of the language production. Two studies look deeper into the potential value-added by formulaic language in NNS speech.

Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006) used a quasi-experimental design to study the effect of formulaic language on experienced EFL teacher's perceptions of fluency and proficiency. Upper intermediate to advanced English majors in Belgium (*n=32*)

were divided into a control (*n=15*) and experimental EFL (*n=17*) class. Over 22 hours of instruction, each group had the same teacher and used the same materials, but the experimental group completed activities to draw their attention to formulaic sequences. These activities included highlighting sequences in a text, cloze activities, and peer work to identify vocabulary patterns. In the control group, the students discussed grammar and single-item vocabulary words. Boers et al. (2006) used a broad operational definition of formulaic language that incorporated formulaic language with different length, transparency, and function. Collocations, idioms, and discourse organizing expressions were all included for consideration.

Data collection included two interviews. In the first interview, the NNSs read an article from a newspaper or magazine and then discussed the article with a test administrator. In the second interview, the NNSs had a spontaneous conversation about an unfamiliar topic. Each interview was scored by four judges. Judge A scored the overall proficiency level of the students, and Judge B assessed the fluency, range of expression, and accuracy of the speech samples. Judge A and B used the same rubric but were responsible for different sub-scales. Judge C and D listened to the interviews and counted the number of formulaic sequences based on their intuition.

Using a Mann-Whitney correlation, Judge A rated the proficiency of the experimental group (*M=14.44*) significantly higher than the control group (*M=13.31*) ($p < .05$). Judge B gave significantly higher ratings ($p < .05$) to the experimental group for fluency (*M=12.53*) compared to the control group (*M=10.60*). The experimental group (*M=12.29*) also outperformed the control group (*M=9.87*) in range of expression scores ($p < .05$). There was, however, no difference in accuracy between the two groups ($p > .05$). Judges C and D found

that the students in the experimental group (*M=11, M=13.57*) used significantly more ($p < .01$) formulaic language in the interviews than the control group (*M=6, M=5.54*).

Using a Spearman rank correlation, the researchers determined if the differences in formulaic language between the two groups contributed to the differences in scores. They found that the formulaic sequence counts of Judge C ($rs = .33;\ p < .05$) and D ($rs = .61;\ p < .01$) correlated with the assessment of overall proficiency. There was also a significant correlation between the formulaic sequence counts of Judge C ($rs = .45, p < .05$) and Judge D ($rs = .44, p < .05$) and the fluency scores. For range of expression, the results were also significant for Judge C ($rs = .39, p < .05$) and Judge D ($rs = .45, p < .05$). The number of formulaic sequences, however, did not significantly correlate with accuracy scores for either judge.

Boers et al. also reported that if the scores for the two interviews were looked at separately, then the experimental group did not produce significantly more formulaic sequences than the control group, and therefore, the greater benefits of formulaic language were mostly confined to the first interview which required discussing a specific article. Furthermore, when the researchers looked more closely at what formulaic sequences were actually used in the first interview, almost one-third of the sequences came directly from the reading passage, not from class discussions of formulaic language. In comparison, only one-fifth of the formulaic language used by the control students came directly from the article. The researchers concluded that formulaic language was beneficial for increasing the perceived fluency of NNSs on structured oral proficiency interviews, but the increase in formulaic language of the experimental group could not be attributed to a larger formulaic knowledge base, even after participating in 22 hours of direct instruction.

Stengers, Boers, Housen, and Eyckmans (2011) replicated Boers et al. (2006) to determine if formulaic language was equally beneficial in improving perceptions of fluency, range of expression, and accuracy for students of different target languages. In this study, high intermediate English majors (n=26) and Spanish majors (n=34) at a Dutch university completed one oral proficiency interview. The interview consisted of reading a 600-word passage in Dutch and then re-telling the story to a teacher in the target language. During the interview, the participants did not have access to the article but were given a list of single-word context clues that corresponded with the chronological events of the story. The researchers used a Dutch article for the input due to the unexpected findings in Boers et al. (2006) where students drew more formulaic language directly from the article than from class instruction.

Native speaking judges of English and Spanish were asked to identify correct, target-like formulaic sequences used in the speech samples. Like Boers et al., the allowable target-like formulaic sequences included strings of various length and semantic transparency. At the same time, different English teachers (n=3) and Spanish teachers (n=3) listened to the recordings and awarded scores for fluency, range of expression, and accuracy using a rubric.

The results indicated that on average English learners (*M=9.77*) used more target-like formulaic sequences than Spanish learners (*M=5.61*). The researchers also found that the average density of formulaic language, or sequences used per minute, differed between the English (*M=2.75)* and Spanish *(M=1.56*) groups. Using a Pearson correlation, these differences were found to be significant ($p < .001$). The English and Spanish teachers' fluency, range of expression, and accuracy scores were averaged into one composite score for each category. The researchers used the formulaic language counts to determine if there was a correlation between

amount of formulaic language and each of the three sub-scale scores. The amount of formulaic language significantly correlated with all three sub-scales for both groups; however, the relationship was stronger for EFL students ($p < .01$) than Spanish as a foreign language students ($p < .05$). The correlation coefficients showed that formulaic language was a predictor of fluency ($r = .55, .36$), range of expression ($r = .63,\ .39$), and accuracy ($r = .56,\ .36$) for English and Spanish language groups respectively. In conclusion, the number of formulaic sequences did contribute to teachers' perceptions of fluency, and the effect was larger for English students.

Studies on the Direct Instruction of Formulaic Language

It is helpful to begin the discussion from a longitudinal perspective. Li and Schmitt (2009) looked at the writing assignments of one graduate student, Amy, over the course of her MATESOL program in England. The purpose of the study was to determine how formulaic sequences are acquired and to document the growth of formulaic knowledge over time. The participant was a 29 year old female from China with 10 years of English language study. In China, Amy worked as a middle school English teacher for four years. Amy's goal was to move up to teaching English at a university. She was classified as an advanced non-native English speaker based on a score of 6.5 on the International English Language Testing System (IELTS).

Li and Schmitt collected nine writing assignments spaced throughout the 10 month graduate program which included eight course-embedded literature reviews and one graduate thesis. These writings created a learner-specific corpus. The researchers also conducted nine semi-structured interviews within one week of completing each writing assignment. All interviews were conducted in Chinese, translated, and transcribed by the bilingual co-researcher.

During the interviews, the researcher showed Amy formulaic sequences taken from each writing sample one by one. The participant indicated how she thought she learned the phrase and her comfort level in using it correctly. Formulaic sequences were not repeated in the interview stage even if they appeared multiple times in her writing.

Li and Schmitt operationalized formulaic sequences based on Nattinger and DeCarrico (1992) who described multi-word units that frequently occur in the same form and function roles to generate idiomaticity. The researchers asked a panel of three native speaking judges to read papers and highlight formulaic sequences. A separate panel then rated the intuition-identified sequences on a five-point Likert scale for appropriateness. In the final stage of the methodology, the study actually became a hybrid case study and longitudinal intervention. The panel's perceptions were shared with Amy, including panel-generated suggestions for more appropriate formulaic sequence in some cases.

The researchers found that, based on a normed frequency count of 700-word sections of writing assignments, Amy used one formulaic sequence every 35 to 49 words. This equates to 2-3% of running words in a text. A review of the panel feedback suggested that, despite the low number, the participant used some sequences too frequently which contributed to the non-native-like quality of her writing (p. 95):

> *Sample 1*- Too much use of discourse markers too close together. They are useful to set out organization of argument, but overuse feels unnatural.
>
> *Sample 5*- I think native-like essays would probably try to use a wider range of forms
>
> *Sample 8*- The only things I notice overall [is]…a tendency to repeat certain phrases a little too often, (e.g. such as).

Li and Schmitt commented that the development in Amy's formulaic knowledge actually hurt her writing as she overused new expressions to such a degree.

Based on the self-report data from the interviews, Amy knew 125 sequences before beginning the MA program and learned 166 during the 10-month period. Sequences acquired from explicit learning occurred from the intensive English program (51), feedback from the judges (15), and the dictionary (2). Academic reading (70), peers (18), and input from spoken language (2) were sources of implicit learning. In this study, a combination of implicit and explicit learning accounted for the largest knowledge gains in formulaic language. Amy's command of the sequences also improved over the 10-month period. The percent of error-free sequences used increased from 40% to 90%. In this study, however, note that command is operationalized as situationally-appropriate, not grammatically correct. Tense and aspect errors continued throughout the course of the study. The participant reported a steady increase in confidence in using acquired formulaic sequences. For example:

> I first noticed [according to] in my friend's writing note, I thought it should be quite useful for my own essay, so I just picked it up…but, after I saw the feedback I realized that I did not use it correctly...because I used "according to somebody's study" where it should be "according to somebody". So now I feel more confident to use this phrase. (p. 97)

Overall, the authors conclude that the participant's growth in formulaic sequences was incremental and not remarkably different from that of traditional vocabulary words. They also note that her use of sequences was "quite conservative" compared to native-like norms.

Schmitt, Dornyei, Adolphs, and Durow (2004) described the acquisition of 20 target formulaic sequences by advanced non-native speakers (*n=94*) to determine if age, gender, language aptitude, and motivation were significant predictors of acquisition. The participants were adults studying at an EAP program in England. The selection criteria for the sequences included frequency, applicability to an EAP curriculum, and pedagogic value. The researchers drew sequences identified by Biber et al. (1999), Nattinger and DeCarrico (1992), and Hyland (2000). The researchers calculated the frequency of the expressions in three comparison corpora of general, general spoken, and academic speaking, or the BNC, CANCODE, and MICASE. The team then searched seven textbooks used in the EAP program to find which sequences were already embedded in the curriculum. Finally, 45 sequences from both stages were included on a survey of EAP instructors to evaluate their usefulness in order to narrow down the target sequences to 20.

Participants completed a productive and receptive test of formulaic language, language learning aptitude, attitude, and motivation assessments, and two measurements of vocabulary size. The productive test was a combination cloze and c-test. The receptive measure was a multiple choice test attached to a reading passage with the target sequences omitted.

Between pre and post-tests, participants' receptive knowledge increased by 12% while productive knowledge grew by 25%. The growth in productive knowledge was significant at the $p<.001$ level on a paired samples t-test. Score on the Vocabulary Levels Test (Laufer & Nation, 1999) only increased by 3% at the 3,000 level and 12% at the 5000 level. It should be noted that the pre-test scores were much higher on the receptive and 3000 level test, so there was less room for improvement. The results of a Wilcoxon paired samples test showed no significant

relationship between productive formulaic knowledge and vocabulary size at either level. The same was true for receptive knowledge. In addition, there was no significant relationship between growth in formulaic language and any of the affective variables.

Jones and Haywood's (2004) descriptive, exploratory design looked at possible instructional approaches that would raise students' awareness, facilitate learning, and increase accuracy in formulaic language. Intermediate ESL students from two intact EAP classrooms participated in the study over the course of ten weeks. One class acted as the experimental group (*n=10*) and the other the control (*n=11*). Jones and Haywood selected 74 three, four, and five-word lexical bundles from Biber et al.'s (1999) study. During their reading class, students were exposed to two types of experimental reading passages. The first increased the frequency of the target bundles and the second bolded the bundles in the text. In addition to traditional reading class assignments, students worked to classify bundles from the readings into functional categories and were exposed to concordance lines from different appropriate uses. The reading class work was carried over into the writing class. Teachers reviewed the sequences and their grammatical and functional roles in discourse, practiced with fill-in-the blank exercises, and compiled examples of use in other writing using concordance lines.

One of the benchmarks for growth involved the analysis of two writing assignments from each student in the treatment group. A panel of EAP teachers identified formulaic language in the writings and gave the phrase a score based on appropriateness. The researchers found no change in the number of formulaic sequences used and in the appropriateness score. In a pre and post c-test of productive skills, five of seven experimental participants increased their scores. During a follow-up interview with three participants, all reported an increased awareness in the

importance of formulaic expressions in writing. Possible reasons for the results include a small window of only two weeks between collecting writing samples and a small sample size.

In a study of native-speaking university students (*n=8*), Cortes (2006) explicitly taught 35 four-word formulaic sequences in a writing-intensive history class. Cortes identified the sequences using a frequency-based corpus analysis. The instructional intervention consisted of five 20-minute lessons presented in two-week intervals during a 10-week class session. Each lesson focused on a group of functionally related bundles. The five functions used in the study were topic elaboration (*on the other hand*), focus, (*one of the major*), quantity specification (*as part of the*), framing attributes (*the extent to which*), and time reference (*by the end of*). Excerpts from history journal articles used in the identification corpus provided examples of sequence use. Dyads analyzed the functions of the expressions in context then completed either a cloze passage, multiple choice, or identified inappropriate sequences in a passage. Cortes collected a pre, mid, and post writing assignment naturally assigned through the course.

After analyzing the papers, Cortes found that 13 of the 35 target sequences were used in the first paper before the intervention. The mid and final writing passage did not show a major increase in the use of target bundles. Only sequences that framed attributes seemed to increase but not to a great extent. Despite informal feedback from students and the course instructor that "students perceived that these lessons had raised their awareness…of lexical bundles in published writing, and that they regarded target bundles as tools that could provide their writing with certain…sophistication…the frequency of their use did not increase significantly." (p. 398). Students continued to opt for single words like conjunctions and adverbs to do the work of sequences.

A Synthesis of Findings- Answers Leading to More Questions

There is a difference in the amount and type of formulaic sequences used in writing, yet status as a non-native speaker might not be the definitive grouping variable for writers that fail to mimic native-norm standards. Experience is key. As writers gain experience, their use of formulaic language begins to meet or get close to meeting the norms established by expert-level writers. As language learners gain experience with the L2, their use of target sequences also increases. These findings are congruent with the connectionist model of SLA (N. Ellis, 1998; 1999; 2003; 2005) in that exposure and practice drive acquisition. What is curious is the issue of syntactical and morphological errors in formulaic sequences produced by non-native speakers if they are indeed stored and recalled in holistic units.

It is difficult to extrapolate the cause of errors in formulaic language production as they are documented in a peripheral fashion in the previously mentioned studies. In fact, Yorio (1989) was the only researcher that specifically questioned why errors should persist in formulaic language used in K's writing. Wray (2002) postulates that errors in formulaic production are not cause to discredit the notion of cognitive cohesiveness. Formulaic language is produced under the same pressures and constraints of a developing interlanguage, and therefore subject to the same developmental and idiosyncratic errors. Consider an Arabic learner of English. Common lexical mistakes for this learner would be negative transfer of spelling rules from the L1 as in *beople* for *people* or universal developmental errors with inflectional morphemes such as *I study in New York for 3 months last year*. The input received certainly included correct examples of how to spell *people* and how to form and when to use the past tense of *study*. The very nature of L2 acquisition is the time it takes to make input and output match, so the same should be true for

extended patterns of vocabulary. Unfortunately, as detailed error analyses of formulaic language are not prevalent, it is impossible to determine if there are differences in the number and type of errors within and outside of sequences.

Another consistent finding is that non-native writing is additionally marked by an overuse of specific sequences. What is more, these repetitive sequences are exclusively of the high frequency variety. This is notable as academic and technical language is characterized largely by lower frequency words. Highly frequent sequences are not the fix-all for fluent academic writing. Salient features identified as having a low frequency but strong association measures or native-speaker norms must also be integrated.

The problem of overuse is further complicated by the nature of the experimental designs investigating the formulaic language gap. It is unclear if the overuse is representative of a fossilized formulaic language or merely indicative of recently noticed sequences. As connectionism calls on practice to strengthen the acquisition of sequences, overuse might be a natural part of the learning curve. Writing teachers note a similar phenomenon anecdotally when a new adjective appears profusely in a writing assignment to create awkward descriptions. Parents note the overgeneralization of syntactical morphemes to produce utterances such as *daddy goed to the store.* If the heteromorphic distributed model equates a sequence with one traditional vocabulary word or morpheme, then overgeneralization would be an expected developmental experience.

Finally, research quantifying the gap in formulaic language continues to draw on larger and larger samples of language. This adds strength to the findings. However, the budding research on instructed formulaic language is limited by small sample sizes and a lack of delayed

post-tests. The results indicate lackluster increases in production from both inductive and deductive explicit instruction. The strongest findings from the research indicates success in increasing students' awareness of bundles. Again, this may in fact show gains in productivity in the long term as noticing is an accepted kick-start to implicit learning processes. At this point, though, it is unclear.

Corpus-Based Lists of Formulaic Language

At the beginning of the chapter, the researcher described corpus-based vocabulary lists for general and academic purposes. These lists, specifically the General Service List (West, 1953) and Academic Word List (Coxhead, 2000), are cornerstones of vocabulary instruction for second language learners. The lists help generate graded readers and proficiency assessments. The popular vocabulary analysis website Lextutor (Cobb, 2014) uses them to analyze the vocabulary in texts and student writing. Corpus-based vocabulary lists continue to be hot topics at teachers' conferences.

For formulaic language, a large number of studies extrapolate salient sequences from a variety of corpora with a variety of tools, yet none of the results have taken hold as a pedagogically-friendly vocabulary list in the same was as single-item lists. Most results, like Biber et al. (1999) are prohibitively long for classroom application with thousands of sequences reported in the findings. Altenburg's (1998) results produce less sequences than Biber and his colleagues, but the list is close to 7,000. Aside from the sheer size of corpus-based analysis, some sequences identified as salient elicit feelings of ambiguity for teachers and material writers. It is unclear how pedagogically important expressions like *is one of the* and *is in a* are

(Altenburg, 1998; Biber et al., 1999). Pedagogically-focused corpus-based formulaic lists are limited. Only two were found in a survey of the literature.

PHRASE List

The Phrasal Expressions List, or PHRASE, consists of 505 formulaic sequences (Martinez & Schmitt, 2012). The researchers used a mixed-methods approach relying on both psycholinguistic, or frequency, and phraseological criteria, or non-compositionality. First, Martinez and Schmitt extracted all two to four-word sequences from the BNC that occurred at least 787 times. Because the BNC organizes single words into word families in 1,000-word frequency bands, the researchers determined that an overall frequency of 787 would meet the cut off range for the 5,000-word band. The first stage of the study produced 15,000 formulaic expressions.

In the second stage, the researchers applied three core criteria to each sequence. The criteria are an extension of the heteromorphic distributed lexical model (Wray, 2002). A sequence had to meet at least one of the three core criteria. First, the expression had to be conceptualized as a morpheme equivalent unit. The second and third criteria relate to an item's compositionality. Martinez evaluated both the transparent and non-transparent semantic qualities. The article provides examples of more transparent expressions like *at this time* and more nebulous expressions such as *for some time* and *every so often.*

In order to make sound judgments for the second and third criteria, supplemental resources were incorporated. First, corpus-informed dictionaries were used to see if a BNC sequence was also listed as a lexical item. The researchers also found that many of the expressions were polysemous. Martinez randomly sampled 100 concordance lines from the

BNC containing the target expression and deleted any occurrence that did not reflect the non-transparent meaning. This test created a percentage (e.g. 100 concordance lines with 30 non-formulaic occurrences equaled 70% formulaicity). These results were verified with a second random sampling to see if the same percentage could be produced. In cases where two samples did not yield consistent results, the procedure was repeated until a reproducible figure could be found. This percentage was then used to readjust the final frequency. For example, the expression *at first* occurred 5,090 times with a formulaic figure of 84%. The final adjusted frequency for the expression was 4,275 (p. 312). Frequency counts were also adjusted to account for inflection. The frequencies of all inflected forms of sequence were compiled into one figure. The study also allowed for open frames such as *shake one's head*. Martinez expanded the frequency counts to include possible forms such *my, your, his*, and *her* and inflected forms of the verb such as *shakes, shaking, shook.*

The final list included 505 sequences that Martinez and Schmitt note are of value in improving L2 learners' receptive vocabulary skills. An interesting secondary finding is that almost all of the sequences are made up words in the 2,000 most frequent single-words of the BNC. In fact, 95% are based on the first 1,000 words. Martinez and Schmitt evaluated a paragraph of 67 words from an academic textbook using Lextutor (Cobb, 2014) which calculates the percentage of the first 1,000 band, second 1,000 band, AWL, and off-list words. The first analysis showed only 7.5%, or 5 words, were off-list. When the same paragraph was analyzed using PHRASE as the identification tool for off-list words, the number of unknown words rose to 26.9%. The authors conclude that PHRASE is a pedagogically salient tool to help reflect the real complexity of texts used in the L2 classroom.

The Academic Formulas List

The Academic Formulas List (AFL) combines corpus linguistic, cognitive science, pedagogical, and English for academic purposes perspectives (Simpson-Vlach & Ellis, 2010) to identify the most common formulaic sequences used in academic language. The AFL is divided into three 200 item sub-lists: core bundles in both academic speaking and writing, academic writing, and academic speaking.

The AFL departs from traditional psycholinguistic, corpus-based studies by including frequency along with mutual information scores and teachers' perceptions of pedagogical value to identify all three, four, and five word formulaic sequences in the corpus. The researchers used a sample corpora of 2.1 million words for academic writing and 2.1 million words for academic speaking. The disciplinary make-up for both corpora included humanities and arts and social sciences. Academic speech also included biological sciences, physical sciences, and a miscellaneous category. Academic writing included natural sciences/medicine and technology and engineering. The researchers used comparison corpora of non-academic speech and writing to validate the resulting list. The AFL also uses a multi-step methodological approach.

From the psycholinguistic perspective, the researchers used frequency, log likelihood, and range criteria. First, the researchers extracted all formulaic sequences that occurred at least ten times per million words in the academic and non-academic corpora. Once overlapping items were deleted, they determined the frequency of each sequence in all four of the sub-corpora. The sub-corpora counts were used to test the relative frequency of each item using the log likelihood. Any sequence with a statistically significant ($p<.01$) relative frequency, or log likelihood statistic, remained for consideration under the range criteria. Common sequences from the

academic speech sub-corpora had to occur in four of the five disciplines while writing sequences needed a range of three out of the four disciplines. In order to be considered on the core list, a sequence needed an overall range of six out of nine disciplines. After applying these three criteria, the number of formulaic sequences was reduced from 2,000 to 979 in speaking, 712 in writing, and 207 core items.

The researchers then integrated MI scores (Oakes, 1998). Simpson-Vlach and Ellis (2010, p. 494) illustrate the need for a balance between frequency and MI with an example from their academic corpus. The expression *blah blah blah* has the highest MI from the spoken academic corpus. This means that it is highly unlikely to hear to word *blah* alone, or even in a set of two. It is almost always used in a sequence of three to gloss over details or as a dismissive. The practical importance of this sequence, however, is limited especially considering an academic speaking environment. In addition, this expression when ranked by frequency is very low. The rationale behind integrating MI was twofold. First, one limitation of corpus-based research is related to high frequency words. In identifying words that commonly occur together, researchers cannot always distinguish between common sequences that result from the natural use of high frequency words and sequences that commonly occur because they are glued together (Schmitt, 2010; Simpson-Vlach & Ellis, 2010; Wray, 2002). In other words, corpus studies can only reveal what language was used, not what could have been used. In addition, there is empirical evidence that native speakers (NS) may be highly sensitive to MI, in fact more so than non-native speakers (NNS) (Durrant & Schmitt, 2009; Ellis, Ellis & Simpson-Vlach, 2009; Simpson-Vlach, & Maynard, 2008).

The researchers created a second list of formulaic sequences ranked by their MI score alone and noted a complete re-ordering of the list (Table 3). The researchers noted that not all of the shifts are helpful. In writing, frequency alone ranked expressions such as *is sufficient to* and *weight of the* at the bottom while it could be argued that they have a higher pragmatic value than *but it is* and *in the united*. MI, comparatively, does help in moving sequences like *to be of*, *as to the*, and *of each of* to the end of list. The researchers argue that both measurements are important, but when used alone, do not provide the whole picture.

Table 2.

Frequency-Based and MI-Based Rankings of Academic Language

Frequency-Based Sequence Rankings		MI-Based Sequence Rankings	
Examples of highest ranked sequences			
Speaking	*Writing*	*Speaking*	*Writing*
this is the	on the other	blah blah blah	due to the fact that
be able to	in the first	trying to figure out	it should be noted
and this is	the other hand	do you want me to	on the other hand the
Examples of lowest ranked sequences			
Speaking	*Writing*	*Speaking*	*Writing*
if you haven't	is sufficient to	okay and the	to the case
so what we're	weight of the	is like the	of each of
as well but	of the relevant	so in the	with which the

Note. Adapted from Simpson-Vlach & Ellis (2010, p. 494-495)

Ellis, Simpson-Vlach, and Maynard (2008) published the results of three correlational experiments validating the combination of frequency and MI. Drawing from the same corpora that were later used to create the AFL and the same unprocessed lexical bundles, the researchers asked native speakers (*n=11*) and non-native speakers (*n=11)* to judge whether or not a particular combination of words was formulaic. The unit of analysis was the time elapsed between the visual of the bundle and participants' decisions. For NS, the number of words in the bundle had a significant positive correlation and the MI score had a significant negative

correlation with judgment time. NNS judgments', however, were significantly predicted by frequency and length of bundle.

In a second study, the researchers recorded the time it took a participant to start to say a bundle after reading the sequence. In the NS *(n=6)* group, number of words and number of phonemes as well as MI were significant predictors of start time. Only the number of phonemes was a significant predictor of total articulation time. For NNSs, frequency again replaced MI as a significant predictor. Finally, NS (*n=18*) and NNS (*n=16*) were evaluated on the priming effects for filling in the last word in a bundle. NS again responded significantly to MI while NNS did not.

Ellis and Simpson-Vlach (2009) also looked at NS (*n=9*) ability to determine if a bundle was semantically congruent in a context. The researchers first showed text with a missing bundle in the middle: *I think you also { } what causes that.* After an initial reading, the participant received the missing bundle and judged if it was correct or not. For native speakers, frequency was not a significant predictor of judgment, and the researchers reported 'marginal significance' for MI (*p=.06*). Finally, a multiple regression was used to predict reading recognition response time based on the number of words in a bundle, frequency, MI, and instructor's judgment scores of teaching worth used in bundle identification. They found instructor judgment to also be a significant predictor of processing speed (*p<.001*).

Returning to the 2010 study, the final step in generating the AFL was to triangulate the frequency and MI scores with teachers' perceptions of formulaic sequences. The researchers used stratified random sampling to select 108 sequences, half from speaking and half from writing, that represent thee, four, and five lengths, low, medium, and high frequency, and low,

medium, and high MI. They asked EAP teachers (*n=20*) to evaluate the 108 items on a five point Likert scale on one of three possible criteria: if the words represented a fixed phrase (*n=6*), if the words had a unified meaning or distinct purpose in communication (*n=8*), and if the words should be taught as a sequence or expression (*n=6*). After a check of inter-rater reliability, the scores from the survey became the independent variable in a multiple regression to determine the amount of variance that can be explained by frequency and MI measures. Both frequency and MI were significant predictors of teacher perceptions; however, MI was a stronger predictor in accounting for 56% of the variance as opposed to 31% by frequency. The researchers used the equation and applied it to the rest of the sequences in order to generate the final AFL core list, academic writing list, and academic speech list.

Conclusion

Nesselhauf and Tschichold (2002) observed that learners' attention is rarely focused on any size of formulaic sequences in teaching materials. In cases where formulaic language was called out in textbooks, Koprowski (2005) found the frequency and range of 822 expressions from three textbooks indicated they have little practical value. He provides examples of collocations and phrases such as *recommend fully* and *on its last feet*. Granger (1998) expresses alarm over the call for more focus on formulaic language in L2 classrooms as "We do not know what to teach, how much to teach, and least of all how to teach" it (p. 159). While corpus research has begun to shed light on what to teach, especially in the newest iterations of pedagogically appropriate formulaic sequence lists, the field is no closer to understanding how to teach formulaic language.

This study further extends the call for empirical research to include an answer to the question why. In the current body of literature, only one study has connected formulaic language in general to writing scores. More research is needed to support the findings of Ohlrogge (2009) by looking closely at specific types of formulaic language. In addition, research is needed to clarify in what way formulaic language influences the perceptions of second language writing. Finally, studies suggest that too little or too much repeated formulaic language can negatively affect writing, but no studies have been able to quantify the effects of different amounts of formulaic language.

CHAPTER 3: RESEARCH METHODOLOGY

In the field of TESOL, a tremendous amount of effort goes into preparing academically-oriented language learners for success in the post-secondary classroom, which often uses writing to assess students' mastery of course content (Ferris, 2009; Knodt, 2006; Sullivan, 2006; White, 2007). The previous chapter contained evidence supporting the connection between vocabulary and successful writing (Guo et al., 2013; Engber, 1993; Grobe, 1981; Santos, 1988). In addition, evidence was presented on the importance of formulaic language in vocabulary knowledge with implications for communicative competence (Harwood, 2002; Howarth, 1998; Wray, 2002). However, non-native speakers consistently lag behind their native-speaking counterparts in formulaic language use; showing only modest gains in production even after direct instruction (Cortes, 2004; Jones & Haywood, 2004; Schmitt et al., 2004). At present, only one study (Ohlrogge, 2009) has directly evaluated the consequences of the gap in formulaic language in non-native writing as measured by score, which is an important indicator of performance.

The primary objective of the current study is to first evaluate the impact of formulaic language on scores given by ESL writing teachers' to non-native academic writing samples. Second, the study investigates whether it is the presence of formulaic language that changes the writing score or if the number of these vocabulary sequences also has an effect. Finally, the study looks at what area of linguistic competence is affected by formulaic language by investigating changes among particular sub-scale scores on a writing rubric. The current chapter presents an overview of the research design as well as a detailed description of the methodology including the research setting, population and sampling procedures, data collection, and instrumentation. The final section includes the research questions.

Research Setting and Population

The research setting for the current study includes Intensive English Programs (IEPs) associated with universities in the southeastern United States. ESL writing teachers are the target population. The researcher arranged to collect data from eight IEPs. The population of ESL teachers with writing experience at these eight schools is 191.

Intensive English Programs

IEPs serve as a necessary transitional setting for most language learners before university study. Many international students that travel to the United States do not speak English as their first language. In order gain admission to an institute of higher education, they must demonstrate their proficiency in English directly through the TOEFL, and indirectly on standardized entrance exams like the SAT for undergraduate or the GRE for graduate students. Despite years of English language study, most academic hopefuls cannot seamlessly transition from the English classroom in their home country to pass the TOEFL nor be successful in a university classroom setting because of the intense linguistic demands. Public and private universities often have an associated IEP on campus. Admission to an IEP is separate from admission to the university itself. The only language-related admission requirement to an IEP is a proficiency test for placement purposes.

The IEPs participating in the present study have different structures in place for session length, number of proficiency levels, and hours of in-class study per week (Table 4). The breakdown of session length is as follows: six-weeks (1), seven-weeks (2), eight-weeks (2), and fourteen-weeks (3). Six IEPs divide their coursework into five proficiency levels. The remaining programs use three and four levels of study. Finally, the hours of English study per week varies among the institutions: 20-hours (1), 21-hours (1), 22-hours (1), 23-hours (3), 24-

hours (1), and 25-hours (1). All the IEP programs use a skill-based curriculum focusing on reading, writing, listening, and speaking. Classes on current affairs, research paper writing, vocabulary, and TOEFL preparation are elective options for intermediate to advanced students.

Table 3.

Breakdown of Participating IEP Characteristics

Institute	Sessions Per Year	Proficiency Levels	Hours Per Week	Size of Faculty	No. of Participants
A	6 7-week	4	23	31	24
B	6 7-week	3	24	23	6
C	6 6-week	5	22	12	10
D	5 8-week	5	23	20	18
E	5 8-week	5	25	7	7
F	3 14-week	5	23	29	7
G	3 14-week	5	21	50	8
H	3 14-week	5	20	30	27

Research Population Determinations

IEP writing teachers were selected as the target population for the study following the rationale of Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) who designed the ESL Composition Profile, which is the grading instrument used in the present study. First, IEP writing instructors have experience teaching and evaluating compositions based on a variety of topics and disciplines, i.e. general academic language. Another focus of an IEP writing instructor is to develop effective academic writing and editing skills in the different rhetorical modes such as analytic, expository, and persuasive writing. Finally, as IEP writing teachers play an integral role in transitioning non-native speakers to the post-secondary classroom, their perceptions of what makes good writing at the post-secondary level will be integrated into their classroom instruction and assessment. Therefore, Jacobs et al. (1981) argue that their perceptions

can be reasonably generalized to represent the writing skills a new student whose L1 is not English would bring with them to the university classroom after leaving an IEP.

The operational definition of an ESL writing instructor included a minimum of one semester of experience teaching writing to adult second language learners. The experience could be from teaching scenarios including but not limited to an IEP, an English for academic purposes program, a college writing center, or a private English language institute. The writing experience did not have to occur at the IEP where they currently taught. This criterion is based on the researcher's experience as an IEP instructor, a writing instructor, and as an ESL teacher educator.

Sampling Procedures

Convenience sampling (Fraenkel & Wallen, 2009) was used in the present study. Convenience sampling is a non-randomized sample that is easily accessible to the researcher and possesses a set of sought-after characteristics. The researcher collected data from 107 ESL teachers. Five participants were excluded from the data pool. The final sample size for the current study includes 102 ESL writing teachers. Detailed information of the sampling procedures are listed below.

Sample Size Determinations

G*Power version 3.1.7 was used to determine an appropriate sample size. In order to conduct a repeated measures multivariate analysis of variance (MANOVA) with 95% power, a .05 significance threshold, and a strong effect size of .20, a sample size of 70 is required. Since the study also controlled for order of grading effects (Spear, 1997), the sample size determination exceeds that of the power analysis.

When teachers grade a series of student papers, the quality of the previous paper can affect the perception of those graded afterwards. The study uses an essay bank of 16 non-native writing samples. The 16 essays included different amounts of formulaic language in a control group followed by low, mid, and high amount groups. Details of the essays and the methodology used to generate them are listed under the instrumentation section later in this chapter.

The study used a within-subjects design, exposing each ESL teacher to all experimental conditions through a combination of four out of 16 essays. In other words, each participant graded one essay from the control, low, mid, and high experimental conditions. Therefore, a re-ordering of the core essay combinations created 96 within-subjects data collection packets. Recruitment of participants proceeded with 96 as the target number of participants.

Recruitment of Participants

Volunteers were recruited by first contacting the directors of the eight IEPs. Permission was obtained from the directors to distribute a call for participation to instructors and collect data. The call for participation included an explanation of the research, eligibility requirements, and a pre-determined date and time for data collection based on the directors' estimate of when the most teachers would be available. The participants did not receive any compensation for participation; however, a meal was provided during the data collection session. Also, a donation to a local non-profit organization was made on behalf of the study participants at the conclusion of the data collection period. The specific data collection steps are outlined in the section below.

Data Collection Procedures

The university's Institutional Review Board approved the following data collection procedure:

1. A room was reserved at each of the IEPs to collect data, and the date, time, and location was distributed to participants via email.
2. Before arriving at the data collection site, testing packets were assembled which included: an instruction sheet, four essays each followed by a copy of the grading instrument, and a demographic survey. Since the study used a within-subjects design, the researcher assigned each participant to one of the 96 data collection packets. The purpose of the assignment was to ensure balanced data collection to avoid having unequal groups in the final data set. In addition, the assignment ensured that participants within each IEP were exposed to all writing samples in the essay bank. Each packet included a unique number that was used in place of participants' names.
3. During the visit to the IEP, instructors were free to begin the testing packet at any time during the reserved time. Data collection packets were placed on desks before participants arrived. Participants were instructed to complete the packet in one sitting and to not share their experience with any other teacher at the IEP until the data collection window was closed. The data collection followed the drop-in approach instead of a formal workshop approach so that instructors could come during their lunch or planning period.
4. Even though the data collection followed a drop-in schedule, it was reasonable to expect teachers to complete the packet in a maximum of one-hour. This is based on the Jacobs et al. (1981) field tests of the ESL Composition Profile that found a teacher could read and grade a 500 word essay with the Profile in roughly ten minutes. The researcher estimated ten minutes to read the instructions, 40 minutes to grade four essays, and 10 minutes to

complete a demographic survey. In practice, the average time needed to complete the packet was 30 minutes.

5. Data from the rubrics and demographic surveys were entered into a data file for analysis with the IBM Statistical Package for Social Sciences program (SPSS). Because the rubric that was used to grade the tests included unequal possible scores for each sub-scale, all sub-scale scores were transformed to a percentage before entering into SPSS. For example, the mechanics section contained a total possible score of 4. A score of 3 would be recorded as 75.

Instrumentation

Target Formulaic Language

The study used a sample of formulaic language as identified by the academic writing sub-list of the Academic Formulas List (Simpson-Vlach & Ellis, 2010). The Academic Formulas List (AFL) identifies the most common and most salient formulaic sequences used in academic language from a corpus of 5 million words. Simpson-Vlach and Ellis used a multiple regression to determine the ranking of sequences on the AFL that accounts for both frequency of occurrence and native speakers' intuitions regarding the strength of a sequence's meaning and form. A detailed description of the methodology used to generate the AFL and the empirical support of the resulting list were discussed in Chapter 2.

The expanded AFL writing sub-list found in the appendix of the 2010 article was used to select target sequences. The list contained 200 formulaic sequences. First, two three-word sequences were excluded for consideration in the present study: *the United Kingdom* and *A and B.* The first was excluded because the current study takes place in the United States, and the sequence might not organically fit into essays written by students. *A and B* was excluded as a

target sequence because of the difficulty in controlling for amount of formulaic language in the writing samples if any two nouns connected by *and* could be counted as formulaic.

The frequency of the remaining 198 AFL sequences was calculated in an independent corpus. The independent corpus was the Michigan Corpus of Upper Level Student Papers, MICUSP. As described in Chapter 2, the corpus does classify native and non-native papers; however, based on personal communication with the corpus director, it was not advisable to use the MICUSP to make comparisons between NS and NNS vocabulary usage. The reasons include an unbalanced sample of NS and NNS writing and unclear wording on the demographic survey accompanying submissions. The surveys did not distinguish between 'what is your L1' and 'what language are you most comfortable communicating in an academic setting.' Therefore, total frequency was used, which is the aggregate of occurrences in both groups included in the corpus. The researcher ranked the 198 sequences from the AFL writing sub-list based on their total frequency of occurrence in the MICUSP.

Operationalizing low, mid, and high amounts of formulaic language. A multi-step process was used to operationalize the amount of formulaic language in the writing samples. Jacobs et al. (1981) estimate that a 30-minute timed writing assignment produces an average composition length of 450 to 500 words for proficient writers. Details of the writing prompts that used in the study are described later in this chapter; however, in short, the prompts were designed to elicit 500-word academic writing samples in a particular time limit. The length of essays was important in operationalizing low, mid, and high amounts of formulaic language.

Corpus-based studies that broke down the proportion of general and academic vocabulary in academic writing were reviewed (Table 5). Nation (2001b) estimated that academic

vocabulary makes up 10% of running words in academic language. Xu and Nation (1984) found that the University Word List (UWL), a pre-cursor to the AWL, accounted for 8.5% of words in the Lancaster-Oslo-Bergen corpus. Coxhead (2000) later found that the UWL covered 9.8% of academic words in the AWL corpus. Coxhead also studied the text coverage of the AWL in samples of business (12%), art (9.3%), law (9.4%), and science (9.1%). Schmitt and Davies (2013) found that the AWL covered 7.2% of running words in the Corpus of Contemporary American English and 6.9% in the academic section of the British National Corpus. Hyland and Tse's (2007) evaluation of the AWL found coverage rates in engineering (11.1%), social sciences (11%), and hard sciences (9.3%) for an average coverage rate of 10.6%. Schmitt and Davies (2013) reported the New Academic Vocabulary List (AVL) covering 13.8% of Corpus of Contemporary Academic English and 13.7% of the academic sub-corpus in the British National Corpus. When these figures are averaged together, the coverage rate of general academic vocabulary in writing is 10.12%, which supports Nation's original estimation of 10%

If general academic vocabulary makes up on average 10% of a language sample, and the AFL does indeed represent sequences of general academic language (Simpson-Vlach & Ellis, 2010), then 10% of a 500-word academic writing sample would reasonably include 50 academic words. This benchmark was used as the mid amount of formulaic language. For the low amount, the percentage of 5%, or 25 words, was arbitrarily selected. For high, 15%, or 75 academic words, was used. This figure was also arbitrary. The researcher experimented with doubling the coverage rate to 20% but found the quantity of words prohibitive for integration in a timed writing exercise. Each of these word counts were then divided by three in order to

determine the number of three-word formulaic sequences needed for each experimental condition.

Under the low formulaic sequence condition, the writing prompts included eight three-word formulaic sequences, or 24 words, which covered 4.8% of running words in a 500 word essay. Writing prompts at the mid condition included the eight bundles from the low condition and eight new bundles for a total of 16 formulaic sequences, or 48 words, for 9.6% coverage. Finally, the high condition included all of the bundles from the mid condition plus nine additional bundles for a total of 25 formulaic sequences, or 75 words, for 15% coverage. A total of 25 three-word sequences were needed for the study. The decision to use three-word bundles only is based on findings in Biber et al. (1999), Cortes (2004), and Hyland (2008) that many four and five-word bundles have a three-word bundle as its base.

Table 4.

Coverage Rates of General Academic Vocabulary

	AWL Corpus	COCA	BNC	Other
University Word List	9.8%	--	--	--
Academic Word List	10%	7.2%	6.9%	10.6%
Academic Vocabulary List	--	13.8%	13.7%	--

Since the current study utilizes the writing sub-list of the AFL as the basis for selecting target formulaic sequences, care was taken to ensure that the entire list was accurately represented among the low, mid, and high conditions described above. The following procedure was used to select the target formulaic sequences.

a) Eliminated all sequences longer than three words from the 198 sequences on the writing sub-list of the AFL.

b) Eliminated bundles that would not be appropriate for use in a timed writing assignment such as *in this figure; see the table*, etc.

c) Identified bundles that contained the same lemmatized vocabulary words and checked the frequency of the expressions in MICUSP, and independent corpus of student academic writing. The lowest frequency bundle was deleted (*be carried out)* while the higher frequency bundle remained (*been carried out).*

d) Identified groups of three-word bundles that were actually part of a longer bundles (*the total number, total number of*). The researcher deleted the bundle with the lower frequency in MICUSP from further consideration.

e) Ranked the remaining 103 sequences from highest to lowest total frequency in MICUSP using the sort function of Microsoft Excel.

f) Created three equal frequency bands based on each item's rank identified as the high band (rank 1-35), mid band (rank 36-71), and the low band (rank 72-103). Note that item 71 was moved to the mid band group because it had the same MICUSP frequency as item 70.

g) Selected every fourth bundle from each band until a group of 25 bundles was formed

In order to assign the bundles to a specific experimental condition, each experimental group needed to be equally representative of the AFL. From the total group of 25 bundles, nine were from the high band (rank 1-35), eight were from the mid band (rank 36-71), and eight were from the low band (rank 72-103). The researcher distributed high, middle, and low ranks from each

band into the experimental conditions. The resulting experimental groups are outlined in Table 6 and Table 7.

Table 5.

Target Bundle Characteristics in MICUSP and AFL

Bundle	MICUSP Rank	MICUSP Raw Frequency	AFL FTW Score	AFL Raw Frequency
it is important	1	244	1.40	92
it is possible	5	166	.77	175
it has been	9	128	.92	168
should not be	13	110	.51	108
depending on the	17	95	.41	62
to do so	21	88	.63	116
does not have	25	81	.46	52
this means that	29	73	.45	77
if they are	33	66	.30	70
it is necessary	36	63	.39	71
in some cases	40	54	.58	68
be noted that	44	49	.45	45
this does not	48	47	.37	59
should also be	52	44	.30	38
insight into the	56	40	.77	34
it is interesting	60	36	.46	38
be explained by	64	33	.58	32
to carry out	72	24	1.22	62
is affected by	76	20	.30	24
allows us to	80	20	.93	32
it follows that	84	18	.55	65
in both cases	88	16	.41	36
same way as	92	13	.58	32
at this stage	96	11	.88	70
over a period	100	8	.66	30

Table 6.

Formulaic Sequences in Low, Mid, and High Experimental Groups

Low Intensity Condition	*Mid Intensity Condition*	*High Intensity Condition*
it is possible	it is possible	it is possible
depending on the	depending on the	depending on the
this means that	this means that	this means that
in some cases	in some cases	in some cases
should also be	should also be	should also be
be explained by	be explained by	be explained by
allows us to	allows us to	allows us to
same way as	same way as	same way as
	it is important	it is important
	should not be	should not be
	does not have	does not have
	is affected by	is affected by
	in both cases	in both cases
	at this stage	at this stage
	it is necessary	it is necessary
	this does not	this does not
		it has been
		to do so
		if they are
		be noted that
		insight into the
		it is interesting
		to carry out
		it follows that
		over a period

Advanced ESL Student Academic Writing Samples

To create the writing samples, prompts were developed based on White's (2007) guide to assessing university student writing. First, the intended task and background of the writers was defined. The goal was to elicit academic writing samples from non-native speaking graduate student participants in a TESOL Ph.D. program. The graduate students in the study all scored

high enough on the TOEFL and the GRE, which includes an analytic writing section, to gain admission to a large research university. The participants also had a minimum of one semester of experience with academic writing for both reflective and assessment purposes in their graduate classes. The conceptualization of academic writing was based on the position statement of the Council of Writing Program Administrators (2008) that incorporates rhetorical, critical thinking and writing, and language convention-based skills, which is congruent with the sub-scales of the ESL Composition Profile (Jacobs et al., 1981) used to rate the writing samples.

Second, the process writers should engage in during the task was outlined. The simulated writing process was that of a 30-minute independent writing task congruent with high-stakes assessments such as the GRE or on-the-job scenarios (Sullivan, 2006; White, 2007). According to White, successful writers in a 30-minute task are able to turn in second or third draft quality writing in a time frame allotted for only one draft.

Participants were asked to write four argumentative essays. In post-secondary classrooms, expository and argumentative writing assignments are the two most common modes for university students (Bloom, 2006). The current study focuses on the argumentative mode. The length requirement of the essay was between 490 and 500 words. The rationale for the essay length is discussed in detail in the subsequent section called “Writing Sample Controls”. Finally, writers were instructed to not use outside resources to complete the essay but instead draw on personal knowledge and experience.

White (2007) states a critical measure of a successful writing prompt will draw upon a large amount of personal knowledge and touch on a topic the participant is interested in writing about.

As all participants were in a TESOL Ph.D. program, they are in the beginning stages of entering the professoriate. As apprentices, there are four threads of knowledge and skill development in the doctoral program: SLA theory, conducting and interpreting research, teaching, and teacher education. These four threads were used to design the writing prompts. Each prompt only included one task as timed essays are often limited to a simple description and analysis (White, 2007). The prompts used keywords for argumentative writing outlined in White's guide:

Prompt 1- Argumentative, Control

When making admission decisions, universities should stop using standardized tests like the GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

Agree or disagree with the statement and provide support for your answer. You do not need to reference any outside sources.

Prompt 2- Argumentative, Low Experimental Condition

It is too difficult to learn a foreign language when you are an adult. The only people who successfully learn another language begin studying as children.

Agree or disagree with the statement and provide support for your answer. You do not need to reference any outside sources.

Prompt 3- Argumentative, Mid Experimental Condition

The most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester.

Agree or disagree with the statement and provide support for your answer. You do not need to reference any outside sources.

Prompt 4- Argumentative, High Experimental Condition

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provides hard evidence. Agree or disagree with the statement and provide support for your answer. You do not need to reference any outside sources.

The four volunteer ESL writers attended one four-hour workshop during which they generated 16 essays for the essay bank. In order to accomplish this, each participant wrote one essay for each prompt. At the beginning of each timed writing exercise, the researcher gave each participant a sheet of paper that included the prompt and the associated list of target sequences, with the exception of the control condition. Each paper was numbered with a code unique to the participant and experimental condition. The participants saved their writing samples to a flash drive provided by the researcher using their unique code. The schedule of the writing workshops was as follows:

Writing Session

a) Explanation of research- 15 minutes

b) Control timed writing prompt 1 - 30 minutes

c) Low level timed writing prompt 2- 45 minutes

d) Mid level timed writing prompt 3 - 60 minutes

e) High level timed writing prompt 4- 75minutes

f) Demographic survey- 15 minutes

Description of writing sample controls. Steps were taken to control for transfer between writing samples, time on task, and essay length. First, no participant was assigned the same essay prompt more than once as there could be dramatic differences between quality of the first and second writing attempt. Each participant composed four argumentative tasks. Also, all participants completed the prompts in the following order to avoid influencing their use of formulaic language: control, low, mid, high.

The participants completed the control writing assignment with a 30-minute time limit. This time limit is congruent with the essays used to design and validate the ESL Composition Profile (Jacobs et al., 1981). Also, it mirrors the time and production constraints for non-native speakers on high stakes assessments such as the analytic writing subsection of the GRE. Finally, 30-minutes is an established length of time for advanced non-native speakers to compose a 450 to 500-word writing sample (Engber, 1995; Guo et al., 2013; Jacobs et al., 1981). However, the list of required sequences attached to the other three writing scenarios make the 30-minute time limit unrealistic.

The researcher conducted two pilot tests (Table 8) to determine an appropriate time limit for writing responses. It should be noted that the pilot tests used different prompts than were included in the final study. However, the pilot and final writing prompts were both 30-minute writing tasks designed to elicit academic writing skills of advanced non-native speakers without the need to reference outside sources.

A native speaker composed a low condition essay in 31 minutes, mid in 35 minutes, and high in 42 minutes. In a second pilot test, a non-native speaker, J, was asked to compose the low, mid, and high level essays and record the time. J is an intermediate-level English learner

who has studied English in the United States for three years through a combination of self-study and six-week ESL courses offered by a university at a local public library. J is a native Korean speaker studying English to pass the Medical Licensing Examination with hopes to begin residency at a U.S. hospital. The participant completed the control prompt within a 30-minute time limit. A time limit was not imposed for the other three exercises but J was asked to record the time. J composed the low condition in two hours, the mid condition in four hours, and the high condition in three hours. J was instructed by the researcher to stop composing the essay if she reached the two-hour mark, but she did not follow this guideline.

In looking at the figures in Table 8, J wrote 136 words in 30 minutes. If J were to keep up that pace, she should have completed 544 words in the 130 minutes taken to complete essay 2, 1,088 in 240 minutes for essay 3, and 816 words in 180 minutes for essay 4. The actual rate of production was 385, 458, and 450 for the last three essays respectively. The vocabulary requirement clearly influenced the completion time.

When deciding the time limits for the writing samples, the effects of time on task and essay length needed to be balanced. Participants were instructed to keep the essay length between 490 and 500 words. The length of the essay was vital to the study as it contributed to the independent variable of percent coverage of formulaic language. At this range, the percent coverage after incorporating all the target formulaic sequences was not altered at the low, mid, or high amounts. While the coverage rates could be maintained by leveling the length requirements by amount of formulaic language, this would introduce the unwanted variable of essay length effects. Essay length has been found to be a predictor of essay score when vocabulary is held constant (Engber, 1995; Ferris, 1994; Guo et al., 1981).

To balance the time-length effects, a 15-minute increase was added to the control time limit for each level of formulaic language. The suggested time limit was 45 minutes for the low condition, 60 minutes for the mid condition, and 75 for the high condition. This is based on the average time it took the native and non-native pilot study participants to complete task two and four, not considering task number three. The advanced non-native speakers were not forced to stop writing after the suggested time limit passed, as an incomplete essay would be regarded as lower quality regardless of the lexical manipulation. The researcher asked participants to record their official start and stop time for writing.

Table 7.

Comparison of Native and Non-Native Writers' Samples from Pilot Study

Essay 1- Control

	Time	Tokens	Type-token ratio	Lexical density	General words	AWL words	AFL phrases	Off list words	T-units	Error-free t-units
NS	29 min.	560	52	56	83.01	9.73	NA	7.26	34	34
NNS	32 min.	136	68	49	91.92	5.15	NA	2.94	9	2

Essay 2- Low Formulaic Language

	Time	Tokens	Type-token ratio	Lexical density	General words	AWL words	AFL phrases	Off list words	T-units	Error-free t-units
NS	31 min.	473	48	56	84.78	8.88	5.07	6.34	34	27
NNS	130 min.	385	51	49	86.52	5.18	7	8.29	9	23

Essay 3- Mid Formulaic Language

	Time	Tokens	Type-token ratio	Lexical density	General words	AWL words	AFL phrases	Off list words	T-units	Error-free t-units
NS	35 min.	489	44	53	84.08	4.49	9.8	11.43	28	27
NNS	240 min.	458	37%	50%	87.36	3.27	11.8	9.37	19	5

Essay 4- High Formulaic Language

	Time	Tokens	Type-token ratio	Lexical Density	General words	AWL words	AFL phrases	Off list words	T-units	Error-free t-units
NS	42 min.	497	47	55	80.24	10.38	15.09	9.38	26	25
NNS	180 min.	450	43	55	84	11.11	17.3	4.89	22	4

The variables of type-token ration, lexical density, general words, AWL words, AFL phrases, and off list words are all listed in percentages. The variables of tokens, t-units, and error-free t-units are listed as numerical data points.

Descriptive analyses used for writing samples. No changes were made to the writing samples with the exception of correcting three spelling errors identified via Microsoft spellcheck. The purpose of selecting graduate students who already passed the GRE for admission to the university was their advanced writing skills. Therefore, grammatical and lexical mistakes were an accurate reflection of advanced ESL writing. Furthermore, any mistakes made in the production of the formulaic sequences served to illuminate typical mistakes advanced non-native speakers might make in their production.

The vocabulary profiler tool from Compleat Lexical Tutor v.6.2 (Cobb, 2014), which is based on the Lexical Frequency Profiler (Laufer & Nation, 1995) and the Range Program (Nation & Coxhead, 2002), was also used. The vocabulary profiler reports the amount of general

vocabulary as a percent of words from the first and second frequency bands of the British National Corpus, the amount of academic vocabulary from the Academic Word List (Coxhead, 2000), and off-list words. Information related to lexical density was also recorded. Lexical density is a measure of content words to total running words. Type-token ratio, or the ratio of unique words to total words, was also documented. These two figures provided a measurement of the lexical complexity of the writing samples (Laufer & Nation, 2005) and were used to establish the lexical similarity between the samples aside from amount of formulaic language.

Finally, in order to make claims about the comparability of the writing samples, the researcher calculated the number of t-units in each writing sample, the number of error-free t-units, and the ratio of error-free to total t-units. A t-unit (Hunt, 1964) measures complexity of a writing sample. A t-unit is classified as a main clause and all subordinate or embedded clauses that generate a complete unit of meaning. Research has established the t-unit as an objective measure of L2 writing (Larsen-Freeman & Strom, 1977; Perkins, 1980). For the purposes of this study, an error-free t-unit was free from both syntactic and semantic errors. A detailed taxonomy of errors was not included because an error analysis was outside the focus of the present study.

ESL Composition Profile

Jacobs et al. (1981) created the ESL Composition Profile to provide teachers and administrators with a tool to objectively assess an academic-oriented ESL student's writing skills. The ESL Composition Profile, henceforth termed Profile, is an analytic rubric with five dimensions: content, organization, vocabulary, language use, and mechanics. The five dimensions are measured on a sub-scale and then compiled into a total score. Each subscale has four score ranges categorized as excellent/very good, good/average, fair/poor, and very poor skill

levels. The Profile sub-scales are described by the authors not as a component of a successful essay but rather as a different lens from which to evaluate the overall effectiveness of communication. In other words, to ensure that "each cog in the machinery of discourse is interacting appropriately with other elements and contributing its fair share to the smooth, efficient operation of the communication process" (Jacobs et al., 1981, p. 32). The Profile, therefore, was selected to extrapolate teacher's perceptions of writing quality in the simulated student essays.

The Profile was selected over other holistic rubrics to avoid a loss of detail. In a holistic rubric, each level contains a cluster of characteristics that the writing sample should possess. For example, the GRE analytic writing rubric contains six levels. The narrow score range of zero to six could obscure differences in perceived writing quality. In addition, as formulaic language includes elements of both the lexical and syntactic systems (Wray, 2002; Halliday, 2004), much valuable information can be gained from a rubric with separate scales for organization, vocabulary, and grammar.

The second reason the Profile was selected is that it was designed to be used by teachers with a variety of experience levels, a minimal amount of pre-use training, and a short amount of time needed to complete the rubric (Jacobs et al.., 1981). In order to use the rubric, raters were instructed to read the writing sample twice. After the first read, raters completed the sub-scales of content and organization based on their first impression. After the second read, raters filled in scores on vocabulary, language use, and mechanics based on how these elements reinforce or alter their initial impression.

The Profile has been evaluated by several validity tests. Jacobs et al. (1981) outline the Profile using five domains of instrument validity. In terms of face validity, or the ability of the Profile to measure what it appears to measure, the researchers state that the sub-scales are reflective of the internal criteria teachers use to evaluate student writing by encompassing big picture items such as content and organization and more discreet skills such as grammar and spelling. Content validity asks if the Profile does evaluate the type of writing students are tasked with in the real classroom. The Profile can be used to evaluate all modes of writing in any subject matter. It is specifically designed, however, to measure timed writing assignments that assess general writing abilities. To measure concurrent validity, the researchers determined the correlation between Profile scores and two other tests of proficiency: the Michigan Battery (*n=599*) and the TOEFL paper-based test (*n=327*). The Profile correlated with the Michigan Battery at .70 on grammar and .67 on vocabulary. The correlation between the Profile and the structure/writing section of the TOEFL was .62. According to Tabachnik and Fidell (2007), a Pearson correlation over .5 indicates a strong relationship. The Profile, therefore, does positively correlate with other measures of language proficiency.

To determine the Profile's ability to predict academic performance outside of the ESL program, Jacobs et al. (1981) looked at non-native speakers (*n=41*) who scored high enough on the Profile to go straight into first year English composition at the university. The researchers found that 40 of the students received a C or higher in the course. Finally, the researchers addressed construct validity by evaluating the Profile scores of students (*n=110*) on a pre and post-writing assignment to measure their growth in writing skills after a semester of writing instruction at the IEP. The researchers reported a significant difference in writing scores

(*t=12.04, df= 109, p<.05*) with an average pre-test score of 59.5 and average post-test score of 70. The researchers further compared graduate (*n=366*) and undergraduate writers' (*n=233*) scores on the Profile under the assumption that graduate students should receive a higher score because they have more academic experience. Results of an independent t-test suggest a statistically significant difference in profile scores (*t=6.89, df= 597, p<.05*) between graduate students (*M= 73.2*) and undergraduates (*M=66.4*).

As an independent source of validity, Astika (1993) evaluated the reliability of the Profile with non-native writers at a university (*n=210*). ESL writers completed an in-house writing assessment consisting of one 30-minute timed writing exercise. Each writing sample was graded by two raters using the Profile. A third rater was called when there was a discrepancy of more than 10 points. The inter-rater reliability based on a Pearson correlation was .82, which is considered a strong correlation (Tabachnick & Fidell, 2007).

Astika also calculated the inter-rater reliability of the content (.83), organization (.82), language use (.85), vocabulary (.84), and mechanics (.73). All five subscales were significantly correlated, *p<.001*. Astika then used the subscales as independent variables in a multiple regression to predict the Profile composite score.

He found that all five sub-scales were significant predictors of composite score. Astika reported the variance covered by each variable. Vocabulary was the largest predictor of composite score, as it accounted for 84% of the change in composite score. Content covered an additional 8% followed by language, organization, and mechanics accounting for a respective 4, 3, and .3%.

The different values between sub-scales supports the different weights assigned to each category. Mechanics only counts for 5 points. Students can earn 30 points for content, 25 for language use, 20 for organization, and 20 for vocabulary. The lower weight for vocabulary is not congruent with the findings, but the findings overall do support the relationship between vocabulary and perceptions of second language writing (Engber, 1995; Guo et al., 2013; Santos, 1988).

Research Questions

Based on a review of the literature, the following research questions were formulated:

1. Are there significant mean differences in composite score on the Profile (Jacobs et al., 1981) between the control and essays that incorporate three different amounts of formulaic language?

2. Are there significant mean differences in content, organization, language, vocabulary, and mechanics sub-scale scores on the Profile (Jacobs et al., 1981) between the control and three amounts of formulaic language?

Data Analysis Procedures

In analyzing the data, the statistical procedures were minimized in order to avoid inflating the Type 1 error (Fraenkel & Wallen, 2009). The current study used a repeated-measures design in which the same teachers contributed to the measurements of the different dependent variables (Field, 2009). To answer the first research question, a repeated-measures ANOVA was used, as it is an appropriate procedure to test for mean differences in one dependent variable (Tabachink & Fidell, 2009). For the second research question, the incorporation of multiple dependent variables required a more powerful test to determine if significant mean differences exist, so a repeated-measures MANOVA was used to check for mean differences among the five Profile

sub-scales (Fraenkel & Wallen, 2009; Tabachink & Fidell, 2009). Following the MANOVA, two post-hoc analyses were used to fully understand the data. Both one-way ANOVAs and a discriminant function analysis were used to determine how the sub-scales differed among the control and three experimental conditions (Field, 2009). The discriminant function analysis was chosen as a complement to the one-way ANOVAs because the Profile sub-scale scores have been shown to have a strong correlation (Jacobs et al., 1981), so the relationship between the dependent variables may have an effect on the robustness of the one-way ANOVAs when determining where the differences in score are found (Tabachnick & Fidell, 2007).

Conclusion

This chapter provided an overview of the research design, participants, research questions, data collection, and data analysis. The next chapter discusses the results from the data collection and analyses in order to answer the two research questions. The final chapter, Chapter Five, discusses both research-based and pedagogical implications along with areas of future study.

CHAPTER 4: RESULTS

This chapter reviews the findings of the present study, which investigated the effect of formulaic language on scores given by ESL writing teachers. The chapter begins with a review of the research questions and research design discussed in Chapter Three. The non-native writers and their associated writing samples used in the current study are also introduced. Next, descriptive statistics related to the ESL writing teachers are presented. Finally, the steps taken to clean the data and associated tests of normality are discussed along with the results from a series of statistical analyses.

Research Questions

The current study uses a quantitative approach to determine how formulaic language, as identified by the Academic Formulas List (Simpson-Vlach & Ellis, 2010) affects ESL writing teachers' perceptions of non-native writing, as determined by score. More specifically, the study looks at the following research questions:

1. Are there significant mean differences in composite score on the ESL Composition Profile (Jacobs et al., 1981) between the control writing samples and writing samples that incorporate three different amounts of formulaic language?
2. Are there significant mean differences in content, organization, language, vocabulary, and mechanics sub-scale scores on the ESL Composition Profile (Jacobs et al., 1981) between the control writing samples and writing samples that incorporate three different amounts of formulaic language?

Description of Non-Native Speaker Writing Samples

In order to uncover a possible connection between formulaic language and writing score, the study used writing samples from advanced non-native writers. As described in Chapter 3, each non-native writer wrote four timed essays that incorporated, zero, eight, 16, and 25 three-word formulaic expressions taken from the Academic Formulas List (Simpson-Vlach & Ellis, 2010) to represent the control, low, mid, and high experimental conditions. The NNSs were asked to compose a response between 490 and 500 words, thereby ensuring that the target expressions represented 5%, 10%, and 15% coverage of the running words in the essays. The maximum amount of time allowed for the control, low, mid, and high writing prompts were 30, 45, 60, and 75 minutes respectively.

The following paragraphs provide a brief background of each participant and a description of their associated writing samples, which are then summarized in Table 10. The descriptions of writing samples include measurements of length; actual time on task; percentages of target formulaic language, general vocabulary, academic vocabulary, and off-list vocabulary; lexical density; type-token ratio; total number of t-units; error-free t-units; and ratio of error-free t-units. The reported figures for general, academic, and off-list vocabulary, lexical density, and type-token ratio were derived via Lextutor (Cobb, 2014). The calculation of formulaic language intensity, t-units, and error free t-units were manually calculated.

As discussed in Chapter 3, t-units (Hunt, 1974) were used as a standardized unit of writing complexity because Larson-Freeman and Strom (1977) and Perkins (1980) established the t-unit as an objective measure of non-native writing skill. T-units are the smallest unit of

sentence-like meaning regardless of punctuation. Errors were operationalized as either syntactic or semantic.

Participant A

Participant A is a native speaker of Chinese. She began studying English in China at age 12. Before beginning the Ph.D. program, she received her master's degree in English language and literature and went on to teach English to university students in her home country. She is currently in her second year of a TESOL Ph.D. program. She did not study at an IEP prior to beginning the Ph.D. program. Based on self-report, she is somewhat comfortable with writing in an academic style for her graduate classes, manuscript submissions, and other tasks.

Control writing sample. Participant A's control writing sample contained 408 running words and was completed in 30 minutes. The essay contained 87.32% general vocabulary, 4.63% academic vocabulary, and 8.05% off-list vocabulary. Participant A incidentally used two target formulaic expressions (*it is possible, should not be*) in her control essay, which covered 1.47% of running words. The lexical density, or ratio of content words to total words, was .52. The type-token ratio, which compares the number of words used only once to the total running words, was .48. Participant A's control writing sample contained 24 t-units, 9 of which were error free. This means that 37.5% of the complete ideas expressed in the essay contained no meaning or form-based errors. The writing sample contained 11 syntactic errors and 11 semantic errors. It is important to note that there is not a direct relationship between the number of t-units and the number of errors as a single t-unit may contain multiple errors. This essay is henceforth referred to as A1.

Low level writing sample. The writing sample contained 497 running words completed in 33 minutes. The essay contained 89.74% general vocabulary, 7.44% academic vocabulary, and 2.82% off-list vocabulary. The eight sequences in this experimental condition covered 4.83% of running words in the text. The lexical density was .54 and the type-token ratio was .42. The essay contained a total of 33 t-units, 16 of which were error free. There were no syntactic or semantic errors in 49% of the t-units. Twenty syntactic errors and eight semantic errors were identified. This essay is referred to as A2.

Mid level writing sample. For this prompt, the participant wrote 493 words in 40 minutes. The essay contained 85.80% general, 9.94% academic, and 4.26% off-list words. The 16 formulaic sequences covered 9.74% of running words in the text. The lexical density was .51 and the type-token ratio was .42. The researcher found a total of 28 t-units. Ten, or 36% of t-units, were error free. The essay contained 15 syntactic and eight semantic errors. This essay is coded as A3.

High level writing sample. Participant A's final writing sample contained 492 words completed in 55 minutes. The essay contained general, academic, and off-list vocabulary of 89.93%, 13.41%, and 3.66% respectively. The 25 sequences covered 15.24% of running words in the text. In addition, the lexical density was .49 and the type-token ratio was .44. The researcher identified a total of 35 t-units. Twenty of the t-units, or 57%, were error free. The essay included 20 syntactic and four semantic errors. This essay is coded as A4.

Participant B

Participant B is a native speaker of Turkish who begin studying English at age 13. She received an undergraduate and master's degree in foreign language education. She worked as an

English teacher in Turkey and Estonia. She studied for one semester at an IEP in Texas. At the time of the study, Participant B had just completed her first semester of a TESOL Ph.D. program. She reported that she is somewhat comfortable with the academic writing demands of an English medium-of-instruction graduate program.

Control writing sample. The control writing sample from Participant B contained 497 words and was completed in 30 minutes. It contained 82.57% general, 7.33% academic, and 10.10% off-list vocabulary. The essay had a lexical density of .50 and type-token ratio of .44. The researcher identified a total of 23 t-units of which 39%, or 9, contained no meaning or form-based errors. Twelve syntactic and nine semantic errors were found. This essay is coded as B1.

Low level writing sample. This writing sample contained 494 words completed in 34 minutes. The essay included 88.86% general, 8.7% academic, and 2.43% off-list vocabulary. The eight formulaic sequences covered 4.86% of the running words in the text. The lexical density was .52. They type-token ratio was .36. The essay contained 27 t-units. Twenty-one t-units were error free, which translates to 78% of the essay. The essay included three syntactic and three semantic errors. This essay is termed B2.

Mid level writing sample. Under this condition, Participant B wrote an essay containing 506 running words in 42 minutes. The vocabulary breakdown includes 87.15% general, 8.7% academic, and 4.15% off-list. The 16 formulaic sequences in this experimental condition covered 9.49% of running words in the text. The essay had a lexical density of .49 and a type-token ratio of .38. The researcher identified 27 t-units. Seventeen, or 63%, were error free. The essay contained 11 syntactic and seven semantic errors. This essay is termed B3.

High level writing sample. The final writing assignment from Participant B included 500 running words completed in 47 minutes. The essay included 81% general, 15.4% academic, and 3.6% off-list vocabulary. The 25 formulaic expressions covered 15% of the running words in the text. The lexical density was .48. The type-token ratio was .38. The researcher found 35 t-units and 17 of them were error free. Therefore, 49% of the complete units of meaning did not contain errors. The essay included 13 syntactical and 10 semantic errors. This essay is referred to as B4.

Participant C

Participant C is a native speaker of Farsi. She began studying English in the sixth grade and also worked as an English tutor from the age of 15. She received her undergraduate degree in English language and literature and her master's in applied linguistics. She worked as an English teacher in Iran at a private institute. At the time of the data collection, Participant C had just completed her first semester in a TESOL Ph.D. program. Based on self-report data, she is somewhat comfortable with academic writing for graduate school.

Control writing sample. In the first timed-writing task, Participant C produced 546 words in 30 minutes. The essay included 84.21% general, 7.8% academic, and 7.99% off-list vocabulary. Participant C also used one target formulaic sequences (*it has been*) representing .005% of her essay. The lexical density was .51. The type-token ratio was .41. The researcher identified 22 t-units, and 36% of them, or 8 t-units, were error free. The essay included 14 syntactic and 2 semantic errors. This essay is referred to as C1.

Low level writing sample. The writing sample included 497 words completed in 39 minutes. The vocabulary breakdown included 88.73% general, 8.05% academic, and 3.22% off-

list. The eight formulaic sequences covered 4.83% of running words in the text. The figures for lexical density and type-token ratio were .56 and .43 respectively. The researcher identified a total of 26 t-units. Nineteen, or 73%, included no errors. Three syntactic and four semantic errors were found. This essay is termed C2.

Mid level writing sample. This sample had 496 running words. It was completed in 44 minutes. The vocabulary was classified as 87.5% general, 9.07% academic, and 3.43% off-list. The 16 target formulaic sequences covered 9.68% of running words. The figures for lexical density and type-token ratio were .51 and .38. The researcher counted 25 t-units including 18 error-free units. Therefore, 72% of the complete ideas in the essay did not include any errors. Six syntactical and one semantic error were found. This essay is termed C3.

High level writing sample. Participant C's final writing sample included 495 words written in 45 minutes. The essay included 84.64% general, 11.52% academic, and 3.84% off-list vocabulary. The 25 target formulaic sequences covered 15.15% of running words. The lexical density was .48. The type-token ratio was .43. The researcher found a total of 28 t-units. Fifteen of them, or 54%, were error free. The errors identified included 15 syntactic and three semantic errors. This essay is referred to as C4.

Participant D

Participant D is a native Spanish speaker from Argentina. She began studying English as a foreign language at age 11. She received an undergraduate degree in Argentina and master's degree in the United Kingdom in English language teaching, and she worked as an English teacher in primary and secondary schools in Argentina. Participant D studied at an IEP for two years before entering an English medium-of-instruction university. At the time of the data

collection, she had just completed her first semester in a TESOL Ph.D. program and self-reports being somewhat comfortable with academic writing for her courses.

Control writing sample. In her first writing assignment, Participant D composed 420 words in 30 minutes. Her control essay included 82.46% general, 11.52% academic, and 6.02% off-list vocabulary. The lexical density was .55. The type-token ratio was .54. The researcher identified a total of 16 t-units in her essay. Nine of them, or 60%, were error free. The essay included seven syntactic and one semantic error. This essay is coded as D1.

Low level writing sample. Participant D's second writing sample had 512 words composed in 40 minutes. The vocabulary breakdown included 83.79% general, 10.16% academic, and 6.05% off-list vocabulary. The eight target formulaic sequences covered 4.69% of the running words in the text. The lexical density and type-token ratio were .48 and .35. The total number of t-units was 21, which included 15 error free complete ideas, or 71%. The writing sample included four syntactic and two semantic errors. The essay is referred to as D2.

Mid level writing sample. At the mid amount of formulaic language, Participant D's writing sample totaled 389 words completed in 54 minutes. The essay included 83.55% general, 12.85% academic, and 3.6% off-list vocabulary. The 16 target formulaic sequences covered 12.34% of running words in the text. The lexical density was .47. The type-token ratio was .46. The writing sample contained 20 t-units. Of these t-units, 70%, or 14, contained no meaning or form-based errors. Five syntactic and three semantic errors were identified. The essay is termed D3.

High level writing sample. The final essay in the essay bank included 510 running words completed in 55 minutes. The vocabulary breaks down to 79.02% general, 17.25%

academic, and 3.73% off-list. The 25 target formulaic sequences covered 14.71% of running words in the text. The lexical density and type-token ratio were .50 and .43 respectively. The essay had 23 t-units, and 61% of them, of 14, were error free. There were seven syntactic and three semantic errors. The essay is coded as D4.

Summary of Writing Samples' Content and Organization

A review of the writing samples showed that each essay addressed the prompt and was written in an academic style. The samples contained a complete response and did not show any evidence of a writer who ran out of time and ended the essay mid stream of consciousness. Each essay was organized in paragraph form and contained a clear introduction, body, and conclusion.

Assembly of Data Collection Packets

The group of 16 essays described above are henceforth termed the essay bank. Each writing sample from the essay bank was randomly assigned to one of four data collection packets. Each packet contained one essay authored by each NNS representing each experimental condition (Table 9). These packets were distributed to ESL writing teachers for data collection. Based on the descriptive figures of the writing samples, all of the reported characteristics fell into the 95% confidence interval. A further check of the Kolmogorov-Smirnov tests indicated that all characteristics except coverage of formulaic language and off-list vocabulary were normally distributed, $p > .05$. As the researcher manipulated the amount of formulaic language, this result is expected. The variables of general, academic, and off-list vocabulary along with lexical density, type-token ratio, and ratio of error-free to total t-units were not controlled by the researcher. Using box plots generated by SPSS, the researcher found no outliers in any of the categories except for off-list vocabulary. Essay B1 was an outlier for the measurement of off-list

vocabulary. A follow-up review of the lexical density and type-token ratio did not show any outliers. Therefore, the uptick in off-list vocabulary, or technical and low frequency words (Nation, 2001b; Xu & Nation, 1984), did not affect the overall lexical profile of essay B1. Each data collection packet was, therefore, treated as containing comparable writing samples.

Table 8.

Essay Assignment to Testing Packets

	Control	Low	Mid	High
Packet 1	A1	B2	C3	D4
Packet 2	B1	A2	D3	C4
Packet 3	D1	C2	A3	B4
Packet 4	C1	D2	B3	A4

Table 9.

Summary of Writing Samples Used in the Study

	Writer	Length	Time	General	Academic	Off-List	Formulaic Language	Lexical Density	Type-Token	T-Units	Error Free	Percent Error Free
Control	A	408	30	87.32	4.63	8.05	1.47**	.52	.48	24	9	37
(1)	B	497	30	82.57	7.33	10.10	--	.50	.44	23	9	39
	C	546	30	84.21	7.80	7.99	.005**	.51	.41	22	8	36
	D	420	30	82.46	11.52	6.02	--	.55	.54	16	9	60
Low	A	497	33	89.74	7.44	2.82	5.43**	.54	.42	33	16	49
(2)	B	494	34	88.86	8.70	2.43	4.86	.52	.36	27	21	78
	C	497	39	88.73	8.05	3.22	5.43**	.56	.43	26	19	73
	D	512	40	83.79	10.16	6.05	4.10*	.48	.35	21	15	71
Mid	A	493	40	85.80	9.94	4.26	9.74	.51	.42	28	10	36
(3)	B	506	42	87.15	8.70	4.15	9.49	.49	.38	27	17	63
	C	496	44	87.50	9.07	3.43	10.28**	.51	.38	25	18	72
	D	389*	54	83.55	12.85	3.60	12.34*	.47	.46	20	14	70
High	A	492	55	89.93	13.41	3.66	15.24	.49	.44	35	20	57
(4)	B	500	47	81.00	15.40	3.60	15.00	.48	.38	35	17	49
	C	495	45	84.64	11.52	3.84	15.15	.48	.43	28	15	54
	D	510	55	79.02	17.25*	3.73	14.71	.50	.43	23	14	61
M		*484.5*	*40.5*	*85.39*	*10.24*	*4.81*	--	*.51*	*.42*	*25.81*	*14.44*	*57*
SD		*41.59*	*9.03*	*3.25*	*3.28*	*2.20*	--	*.03*	*.05*	*5.29*	*4.23*	*14*

Note. The figures for general, academic, and off-list represent percentages of vocabulary based on categories produced by Lextutor (Cobb, 2013). Formulaic language represents the percentage of the writing sample covered by the target formulaic sequences. Lexical density is a ratio of content words to total running words. Type-token is a ratio of unique words used only once to total running words. These figures were also generated by Lextutor (Cobb, 2013). T-units are the smallest complete unit of meaning (Hunt, 1964). An error free t-unit is free from semantic or syntactic errors.

Description of ESL Writing Teachers

Data collection occurred at eight IEPs in the southeastern United States. The population of ESL teachers with at least one semester of experience teaching ESL writing from these eight institutions is 191. From that population, 107 ESL writing teachers volunteered to participate. After data collection, five participants were eliminated from the data pool for completing the packet incorrectly. The breakdown of participants by institution can be seen in Table 11. Information regarding the teacher demographics and teaching experience are summarized in Tables 12 and 13.

Based on survey data, the majority of the participants were females between the ages of 31 and 40. Several L1s were represented in the sample including English, Spanish, Italian, and Russian. In terms of overall classroom experience, most instructors reported more than five years of ESL teaching experience. For ESL writing experience in particular, only a few of the instructors were currently teaching their first semester of ESL writing. Most had taught ESL writing more than six semesters. About half of the teachers taught a range of proficiency levels, but some identified with a particular proficiency level.

Table 10.

Number of Participants by Institution

Institution	No. of Participants	Percent of Sample
A	22	22
B	6	6
C	10	10
D	18	18
E	7	7
F	6	6
G	8	8
H	25	25
Total	*102*	*100*

Table 11.

Participant Teaching Experience

		Number	Percent
Experience Teaching ESL	< 1 year	3	3
	1-3 years	14	14
	3-5 years	16	16
	> 5 years	62	61
	No answer	7	6
Experience Teaching ESL Writing	First semester	4	4
	2-6 semesters	21	21
	> 6 semesters	70	69
	No answer	7	6
Proficiency Level Typically Taught	Beginning	7	7
	Intermediate	19	19
	Advanced	18	18
	All levels	50	49
	No answer	8	9

Table 12.

Participant Demographics

		Number	Percent of Sample
Gender	Male	32	31
	Female	62	61
	No answer	8	8
Age	20-30 years old	19	17
	31-40 years old	28	28
	41-50 years old	22	22
	50 and older	24	24
	No answer	9	9
Native Language	English	84	82
	Spanish	5	5
	Italian	2	2
	Russian	2	2
	Other	5	5
	No answer	4	4

Testing Statistical Assumptions

Before analyzing the data, a test of the statistical assumptions of normality and homogeneity of variance was needed for each of the variables. There were 24 dependent variables in the study. They included the ESL Composition Profile (Jacobs et al., 1981) composite score and the subscale scores of content, organization, vocabulary, language use, and mechanics for the control, low, mid, and high experimental conditions. All the dependent variables were continuous level data.

First, the dependent variables were checked for outliers using boxplots, which identify outliers using the interquartile range. Any scores greater than 1.5 times the third quartile or 1.5 times smaller than the first quartile are marked as outliers. The initial analysis identified 32 cases where a participant's score on one of the 24 variables was an outlier. A closer look at these cases revealed that nine participants were responsible for 75% of the outlying data points.

The skewness and kurtosis values and Kolmogorov-Smirnov tests (K-S test) were evaluated for the 24 variables including all data and then again by removing the nine multivariate outliers. When all data were included, the skewness, kurtosis, and K-S test indicated that the data were not normally distributed (Table 14 and 15). A re-analysis of the data excluding the multivariate outliers also indicated non-normality with the exception of the control composite and mid composite variables. However, according to Tabachnick and Fidell (2007), the skewness, kurtosis, and K-S test can be unreliable for large data sets with over 100 cases. Therefore, histograms and Q-Q plots were used to visually assess the normality of the data, but the analysis of normality was inconclusive. The results indicated that, overall, the 24 variables were negatively skewed, indicating that the majority of teachers gave the essays higher scores on

the ESL Composition Profile, so only few lower scores were present to complete the bell curve. The data was not transformed to normalize the distribution as Glass et al. (1972) warns that "the payoff of normalizing transformations in terms of more valid probability statements is low, and they are seldom worth the effort (p. 241).

The assumption of homogeneity of variances was also run, which in a repeated measures design is interpreted through a test of sphericity. Sphericity ensures that the variation between the differences of each experimental condition is roughly equal (Field, 2009). Based on Mauchly's test, the data does not violate the assumption of sphericity, $x^2(5) = 3.76, p > .05$.

Table 13.

Tests of Normality for ESL Composition Profile Sub-scale Scores

	All Data Points			Multivariate Outliers Excluded		
	Skewness	*Kurtosis*	*K-S Test*	*Skewness*	*Kurtosis*	*K-S Test*
Control						
Content	-.43	-.47	$p < .001$	-.49	-.36	$p < .001$
Organization	-.89	.88	$p < .001$	-.82	.89	$p < .001$
Vocabulary	-.60	-.05	$p < .001$	-.66	.04	$p < .001$
Language	-.26	-.90	$p < .001$	-.26	-.79	$p < .001$
Mechanics	-.49	-.62	$p < .001$	-.41	-.73	$p < .001$
Low						
Content	-1.03	.67	$p < .001$	-.93	.38	$p < .001$
Organization	-1.07	.52	$p < .001$	-1.29	1.63	$p < .001$
Vocabulary	-.77	.73	$p < .001$	-.65	.88	$p < .001$
Language	-1.17	1.41	$p < .001$	-1.15	1.65	$p < .001$
Mechanics	-.69	-.47	$p < .001$	-.71	-.52	$p < .001$
Mid						
Content	-.60	-.19	$p < .001$	-.52	-.23	$p < .001$
Organization	-.49	-.55	$p < .001$	-.31	-.93	$p < .001$
Vocabulary	-.99	1.78	$p < .001$	-1.15	2.46	$p < .001$
Language	-.52	-.10	$p < .001$	-.29	-.55	$p < .001$
Mechanics	-.32	-.65	$p < .001$	-.31	-.62	$p < .001$
High						
Content	-.55	-.32	$p < .001$	-.55	-.25	$p < .001$
Organization	-1.14	1.86	$p < .001$	-1.22	2.78	$p < .001$
	Skewness	*Kurtosis*	*K-S Test*	*Skewness*	*Kurtosis*	*K-S Test*
Vocabulary	-.39	-.77	$p < .001$	-.46	-.63	$p < .001$
Language	-.79	.37	$p < .001$	-.80	.64	$p < .001$
Mechanics	-.55	-.59	$p < .001$	-.58	-.57	$p < .001$

Table 14.

Tests of Normality for ESL Composition Profile Composite Score

	All Data Points			Multivariate Outliers Excluded		
	Skewness	*Kurtosis*	*K-S Test*	*Skewness*	*Kurtosis*	*K-S Test*
Control	-.29	-.58	$p > .05$	-.30	-.57	$p > .05$
Low	-.80	.50	$p < .001$	-.74	.15	$p < .001$
Mid	-.45	-.59	$p < .001$	-.32	-.67	$p > .05$
High	-.71	.05	$p < .001$	-.72	.22	$p < .001$

Research Question 1

The first research question concerned the effect of differing amounts of formulaic language on the ESL Composition Profile (Jacobs et al., 1981) composite score. In order to answer this question, a repeated measures ANOVA was run.

Main Analysis

The results show that the composite scores on the ESL Composition Profile (Jacobs et al., 1981) were significantly different, $F(3, 303)=6.05$, $p < .001$, $\eta^2 = .06$ (Table 16). The 6% change in composite score between the three amounts of formulaic language is a medium effect size using the conventional interpretation of .01, .06, and .14 as small, medium and large respectively (Field, 2009). Pairwise comparisons (Table 17) showed that the control and three experimental conditions formed two scoring groups. First, there was no significant difference in composite score between the control (M=85.30, s=8.18) and the low writing samples (M=85.67, s=8.81). There was also no significant difference in composite score between writing samples that included the mid (M=81.92, s=9.79) and the high (M=82.48, s=9.89) amounts of formulaic language. The difference between these two scoring groups was, however, significant (Figure 2). The essays in the control group and those that incorporated eight target formulaic sequences outscored the essays that integrated 16 and 25 sequences.

Table 15.

Repeated Measures ANOVA Tests of Within-Subjects Effects for All Data Points

		Sum of Squares	*df*	Mean Square	F	*Sig.*	η^2	Observed Power
Formulaic Language	Sphericity Assumed	1119.24	3	373.08	6.05	.0001	.056	.96
Error	Sphericity Assumed	18696.51	303	61.71				

Table 16.

Mean Composite Scores on ESL Composition Profile

	All Data Points		Multivariate Outliers Deleted	
	M	*s*	*M*	*s*
Control	85.30	8.18	85.73	7.97
Low	**85.66**	8.82	**86.85**	7.91
Mid	81.92	9.79	82.73	8.94
High	82.48	9.80	83.59	9.03

Supporting Analyses

Because initial tests of statistical assumptions indicated that the data might not be normally distributed, the repeated measures ANOVA was re-run after eliminating the nine multivariate outliers. The results of the second repeated measures ANOVA were congruent with the first. There was a significant difference between composite scores for the control and experimental conditions, $F(3, 276)=5.94$, $p<.01$ (Table 18). In this test as well, 6% of the variance in composite score was accounted for by the amount of formulaic language. Pairwise comparisons showed no significant differences between the control (M=85.73, s=7.97) and low (M=86.84, s=7.91) experimental conditions as well as the mid (M=82.73, s=8.94) and high (M=83.59, s=9.03) conditions. The difference in composite score between these two groups continued to be significantly different, as the control and low level essays outperformed the mid and high conditions (Figure 4).

As a final test of fidelity of results due to potential effects of non-normality, the non-parametric equivalent of a repeated measures ANOVA, a Friedman's ANOVA, was evaluated. Friedman's ANOVA is robust when assumptions have been violated and is an alternative approach to determining differences between several related groups (Field, 2009). When all

cases were included, the composite score did significantly change between the control and experimental conditions, $x^2(3) = 20.76, p < .001$. Based on the results of the parametric and non-parametric analyses, the researcher is confident that the violation of assumption of normality did not adversely affect the results.

Table 17.

Repeated Measures ANOVA Tests of Within-Subjects Effects Excluding Multivariate Outliers

		Sum of Squares	*df*	Mean Square	F	*Sig.*	η^2	Observed Power
Formulaic Language	Sphericity Assumed	1003.11	3	33.37	5.94	.001	.061	.96
Error	Sphericity Assumed	15524.89	276	56.25				

Summary of Results

In sum, there was a statistically significant difference in composite score among the essays with manipulated amounts of formulaic language. The essays with the low level of formulaic language, or eight target expressions, received the highest composite scores overall, yet the difference was not large enough to be significantly better than the control. The essays with 16 and 25 formulaic sequences did not have an advantage when scored by ESL writing teachers.

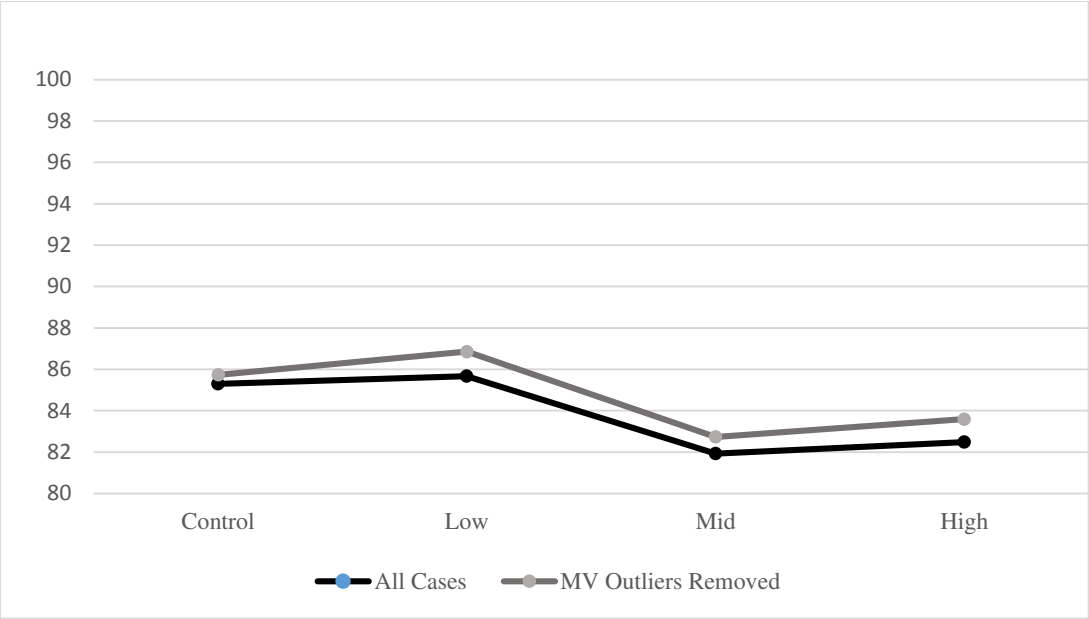


Figure 3. Mean Composite Scores for Experimental Conditions

Research Question 2

The second research question looked further into the scores to determine where potential differences are generated among the five sub-scales on the ESL Composition Profile (Jacobs et al., 1981) when formulaic language is manipulated. The goal was to identify any significant differences in subscale scores, especially in cases that may have been masked by non-significant differences in composite scores.

Main Analysis

In order to determine the effect of formulaic language on the five sub-scales of the ESL Composition Profile (Jacobs et al., 1981), namely content, organization, vocabulary, language, and mechanics, a repeated measures MANOVA was conducted using all cases. All cases were used because the multiple analyses conducted to credit the validity of the repeated measures ANOVA indicated that the results were not affected by the nine multivariate outliers or potential

non-normality. However, when interpreting the results of the MANOVA, the Pillai-Bartlett trace was used as it is the most robust to violations of statistical assumptions when sample sizes are equal (Field, 2009). The results showed that there was a significant effect of formulaic language on the five sub-scale scores, $F(15, 903) = 2.04, p < .01, \eta^2 = .03$. The amount of formulaic language accounted for 3% of the variance in sub-scale scores (Table 19), which is a small effect size (Field, 2009).

Table 18.

Repeated Measures MANOVA Test of Within-Subjects Effects

	Value	*F*	*df*	Error *df*	Sig.	η^2	Observed Power
Pillai's Trace	.10	2.04	15	903	.011	.033	.967

Post-Hoc Analyses

The separate univariate ANOVAs produced as a post-hoc analysis of the main effect (Table 20) showed significant change in content, $F(3, 303) = 3.53, p < .05, \eta^2 = .03$, organization, $F(3, 303) = 4.91, p < .05, \eta^2 = .91$, language, $F(3, 303) = 3.71, p < .05, \eta^2 = .80$, and vocabulary scores, $F(2.87, 303) = 6.85, p < .001, \eta^2 = .98$. The mechanics sub-scale showed a non-significant change between the control and three amounts of formulaic language, $F(3,303) = 2.49, p > .05$. Based on the effect sizes, the organization, language, and vocabulary sub-scales were strongly affected by the manipulation in formulaic language.

Table 19

Repeated Measures MANOVA Univariate Analyses

	Sum of Squares	*df*	Mean Square	F	*Sig.*	η^2	Observed Power
Content	2871.52	3	957.17	3.53	.015	.034	.78
Organization	3335.23	3	1111.74	4.91	.046	.908	.91
Vocabulary	3111.66	2.87	1085.19	6.85	.000	.977	.97
Language	1361.61	3	453.87	3.71	.012	.803	.80
Mechanics	1377.15	3	1361.61	2.49	.061	.614	.61

Note. Mauchly's test indicates that the vocabulary subscale variable violated the assumption of sphericity $x^2(5) = 13.61, p < .05$. The table includes that Huynh-Feldt correction univariate statistics, which is the appropriate correction when Mauchly's W is greater than .75 (Field, 2009).

Content sub-scale. A review of the pairwise comparisons for the control sub-scale showed that the low condition had the highest score followed by control, high, and mid (Table 21). Once again, two scoring groups emerged in the data. The content scores for the control (*M*=75.77*, s*=17.18) and low (*M*=76.69, *s*=17.54) essays were significantly larger than those for the mid (*M*=70.81, *s*=18.26) and high (*M*=71.43, *s*=18.37) experimental conditions. While the addition of eight target sequences increase the content score, it was not a large enough increase to be significant. The addition of 16 and 25 sequences did significantly lower the content sub-scale score.

Table 20.

Pairwise Comparisons of Content Scores across Experimental Conditions

	M	*s*
Control	75.77	17.18
Low	**76.69**	17.54
Mid	70.81	18.26
High	71.13	18.37

Organization sub-scale. This sub-scale followed the same score ranking pattern of low, control, high, and mid (Table 22). The organization sub-scale scores were not significantly different for the control (*M*=79.80, *s*=14.64) and low condition (*M*=80.79, *s*=15.85). There were also non-significant differences between the mid (*M*=74.09, *s*=17.55) and high (*M*=75.28, *s*=17.77) experimental conditions, yet the first two conditions continued to significantly out-score the mid and high conditions.

Table 21.

Pairwise Comparisons of Organization Scores across Experimental Conditions

	M	*s*
Control	79.80	14.64
Low	**80.79**	15.85
Mid	74.09	17.55
High	75.28	17.77

Vocabulary sub-scale. The vocabulary sub-scale also held to the low, control, high, mid scoring pattern (Table 23). However, it did produce slightly different pairwise comparisons. In this case, the control vocabulary sub-scale score (*M*=80.53, *s*=12.48) was only significantly higher than the mid experimental level (*M*=75.57, *s*=16.10). The low vocabulary sub-scale score (*M*=82.91, *s*=11.11) was significantly higher than the mid and the high (*M*=77.86, *s*=14.22) levels. There were no significant differences between control and low, control and high, and mid and high amounts of formulaic language.

Table 22.

Pairwise Comparisons of Vocabulary Scores across Experimental Conditions

	M	*SD*
Control	80.53	12.48
Low	**82.91**	11.11
Mid	75.57	16.10
High	77.86	14.22

Language sub-scale. The language sub-scale is the only variable that produced a different rank order (Table 24). For this sub-scale, control and low scores tied followed by the high and mid conditions. From the pairwise comparisons, control (*M=81.89, s=11.44*) was significantly higher than the mid (*M=77.84, s=13.13*) as well as the high (*M=78.75, s=13.27*) amounts. Essays that included the low level of formulaic language manipulation (*M=81.79, s=17.54*) also received significantly higher language scores than the mid and high level essays. There continued to be no significant difference in score between control and low and mid and high essays in language sub-scale scores.

Table 23.

Pairwise Comparisons of Language Scores across Experimental Conditions

	M	*s*
Control	**81.98**	17.18
Low	81.79	17.54
Mid	77.84	18.26
High	78.75	18.37

Mechanics sub-scale. The mechanics sub-scale scores returned to the scoring pattern of low, control, mid, then high (Table 25). The pairwise comparisons, however, only revealed one significant difference. The low level essays (*M*=87.25, *s*=14.82) scored significantly higher on

this sub-scale than the mid level essays (*M*=82.11, *s*=15.88). No other significant differences were reported between the control (*M*=84.80, *s*=15.40), low, and high (*M*=85.29, *s*=15.48) experimental conditions.

Table 24.

Pairwise Comparisons of Mechanic Scores across Experimental Conditions

	M	*s*
Control	84.80	15.40
Low	**87.25**	14.82
Mid	82.11	15.88
High	85.29	15.48

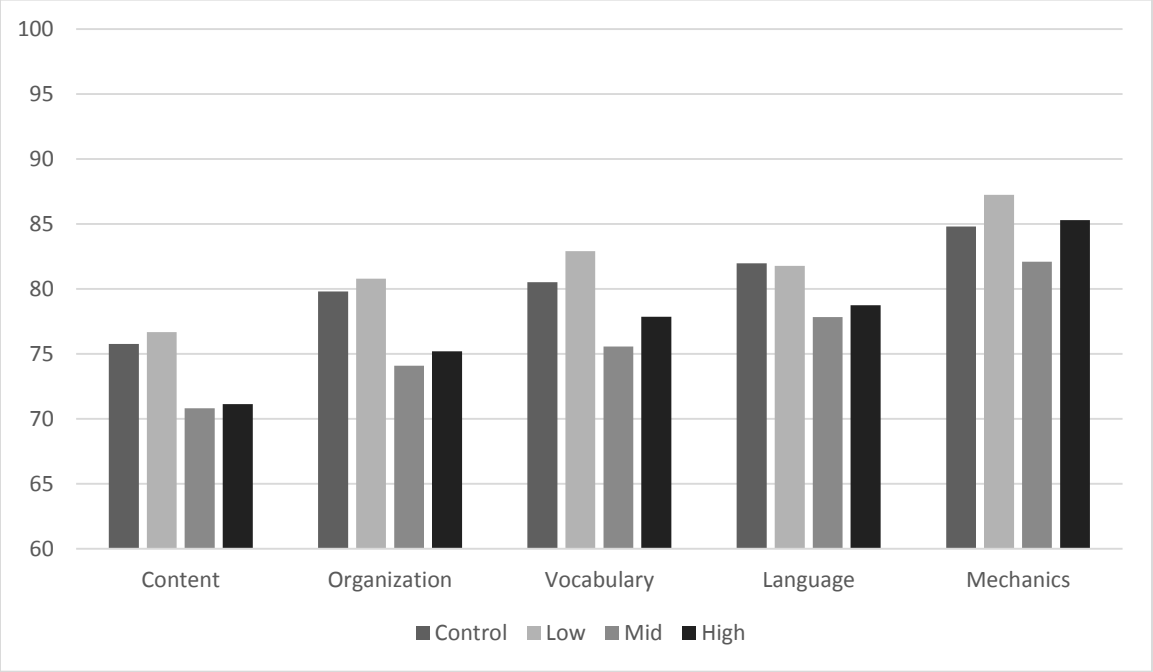


Figure 4. Mean Sub-Scale Scores across Experimental Conditions

Summary of Results

There were significant differences between the sub-scale scores, but the sub-scale scores decreased as the amount of formulaic language increased (Figure 5), thus suggesting an inverse relationship. The essays with the lowest amount of target formulaic sequences continued to score the highest except on the mechanics and language sub-scales. For these sub-scales, the control essays were the highest. However, the difference in score was not significant, and in fact the difference between the control and the low condition was never significant. The mid level essays continued to be the lowest scoring writing samples across sub-scales. The manipulation of formulaic language had the largest effect on organization, vocabulary, and language scores.

Additional Findings

The data were analyzed to determine the comparability of the testing packets since the study used authentic writing samples where the type and number of errors were unaltered. Therefore, it was important to determine if the t-unit error proportion was acting as a hidden latent variable affecting the strength of the independent variable. A mixed effects ANOVA was run to assess the impact of testing packet on the composite score of the essays. The main effect indicated that testing packet did not have a significant effect on composite score, $F(3, 98) = 1.31, p > .05$. However, the test of within-subjects effects produced a significant interaction between formulaic language and testing packet on composite score, $F(9, 294) = 6.48, p < .001$ (See Figure 4). The post-hoc pairwise comparisons were reviewed, and they supported the findings of the ANOVA's main effect. No testing packet was found to be significantly different from the other. Due to these conflicting findings, however, four essays were selected using a

strict interpretation of comparable t-unit error rates and re-ran the analyses from the first and second research questions.

Selecting Essays Using the Strict Error Interpretation

The researcher identified all the essays that contained a t-unit error proportion within one standard deviation from the mean (*M*=.57, *s*=.14) (Table 26). The following essays from the essay bank contain a t-unit ratio between .43 and .71. The actual error-free ratio is shown in parenthesis:

Packet 1- D4 (.61)

Packet 2- A2 (.49), D3 (.70), C4 (.54)

Packet 3- D1 (.60), B4 (.49)

Packet 4- D2 (.71), B3 (.63), A4 (.57)

Next, one essay from each packet was selected. The priority was to choose essays based on packet not NNS authorship in order to only include each ESL writing teacher participant one time in the additional analyses. For example, essay A4 had an error proportion of .57, which is equal to the mean. However, essay D4 was the only essay from packet 1 that met the inclusion criteria. If D4 was eliminated from the analysis, ESL writing teachers who scored packet 1 (*n=24*) would be excluded from the analyses while teachers who scored packet 4 (*n=24*) would be counted twice in the analyses. The final cohesive group of four essays based on t-unit error proportion included:

- Control- D1 with an error rate of .60 from packet 3
- Low- A2 with an error rate of .49 from packet 2
- Mid- B3 with an error rate of .63 from packet 4

- High- D4 with an error rate of .61 from packet 1

The scores for these four essays were compiled into a separate database, which included scores from all participants (n=102), so the demographics listed earlier in the chapter remain the same.

Table 25.

Proportion of Error-Free Units of Meaning by Testing Packet

	Control	**Mid**	**Low**	**High**	***Average***
Packet 1	37%	78%	72%	61%	*62%*
Packet 2	39%	49%	70%	54%	*53%*
Packet 3	60%	73%	36%	49%	*55%*
Packet 4	36%	71%	63%	57%	*57%*

Analysis of Composite Scores Based on Strict Error Interpretation

A one-way independent ANOVA was run to determine if formulaic language affected composite score for writing samples when a strict limitation on the percentage of errors was applied. There was a significant effect of formulaic language on composite score, $F(3, 98) = 2.72, p = .05, \eta^2 = .28$ (Table 27). Using Tukey's test, the multiple comparisons showed only one significant difference in composite score between the control (*M*=88.08, *s*=8.76) and the mid (*M*=81.13, *s*=9.67) essays (Table 28). The addition of eight and 25 target formulaic expressions did lower the composite score but not by a significant amount. Under a strict interpretation of comparable t-unit ratios, formulaic language did not give writers an advantage (Figure 6).

Table 26.

Independent ANOVA for Composite Scores Using Strict Interpretation of Comparable T-Unit Ratio

	Sum of Squares	*df*	Mean Square	F	*Sig.*
Between	630.53	3	201.18	2.72	.05
Within	7582.49	98	77.37		
Total	8213.02	101			

Table 27.

Mean Composite Scores on Profile Using Strict Interpretation of Comparable T-Unit Ratio

	Essay	M	SD
Control	D1	88.08	8.76
Low	A2	83.90	7.36
Mid	B3	81.13	9.67
High	D4	83.17	9.50

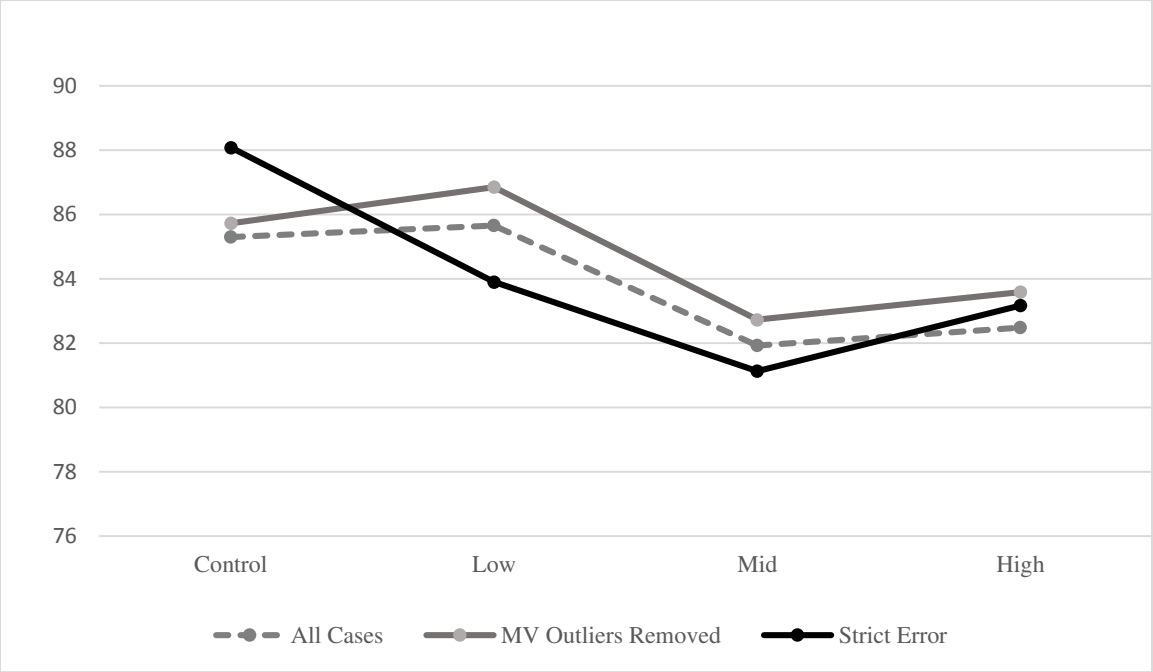


Figure 5. Comparing Mean Composite Scores with a Strict Error Interpretation

Analysis of Sub-Scale Scores Based on Strict Error Interpretation

A MANOVA was also re-run using the data from the four essays to determine if there was a significant difference among the sub-scales under the strict error condition. Using Pillai's trace, there was a significant difference between the control and experimental conditions, $F(15, 288) = 3.85, p < .001$ (Table 29). The amount of formulaic language accounted for 17% of the variance in sub-scale score. However, separate univariate ANOVAs for the dependent variables (Table 30) revealed non-significant formulaic language effects on organization, $F(3, 98) = 1.07, p > .05$ and mechanics, $F\ (3,98) = .34, p > .05$. The content, $F(3, 98) = 3.30, p < .05, \eta^2 = .09$, vocabulary, $F(3, 98) = 9.59, p < .001, \eta^2 = .23$, and language, $F(3, 98) = 3.84, p < .05, \eta^2 = .11$, sub-scales were significantly different between experimental conditions.

Table 28.

MANOVA Results Using Strict Interpretation of Comparable T-Unit Error Ratio

	Value	F	*df*	Error *df*	Sig.	η^2	Observed Power
Pillai's Trace	.50	3.85	15	288	.000	.167	1.00

Table 29.

MANOVA Tests of Between-Subjects Effects for Strict Interpretation of Comparable T-Unit Error Ratio

	Dependent Variable	Sum of Squares	*df*	Mean Square	F	Sig.	η^2	Observed Power
Fom.Lang.	Content	2872.33	3	957.44	3.30	.024	.092	.738
	Organization	900.59	3	300.20	1.07	.367	.032	.281
	Vocabulary	4609.24	3	1536.41	9.59	.000	.227	.997
	Language	1782.05	3	594.02	3.84	.012	.105	.806
	Mechanics	215.04	3	71.68	.341	.769	.010	.114
Error	Content	28417.10	98	289.97				
	Organization	27595.30	98	281.59				
	Vocabulary	15703.51	98	160.24				
	Language	15175.36	98	154.85				
	Mechanics	20624.43	98	210.45				

Post-Hoc Multiple Comparisons for Content, Language, and Vocabulary Sub-Scales

A post-hoc Tukey procedure was reviewed for the three sub-scales with significant differences in the univariate ANOVAs (Table 31). In terms of content, the post-hoc conflicted with the test of between-subjects effects. There were no significant differences between any of the experimental conditions. On the language sub-scale, a post-hoc Tukey only indicated a significant difference between the control (*M*=85.33, *s*=11.08) and the low (*M*=74.71, *s*=13.95) experimental condition. There were no significant differences between any other conditions. Finally, vocabulary showed the most significant effects for formulaic language. There were significant differences between control (*M*=88.29, *s*=7.96) and low (*M*=78.08, *s*=11.60), control and mid *(M*=69.94, *s*=15.04), and mid and high (*M*=83.86, *s*=15.06) vocabulary sub-scale scores.

Table 30.

Multiple Comparisons of Content, Language, and Vocabulary Sub-Scales across Experimental Conditions Using Strict Interpretation of Comparable T-Unit Ratio

		M	*s*
Control	Content	79.56	19.75
	Language	85.33	11.08
	Vocabulary	88.29	7.96
Low	Content	77.40	12.59
	Language	74.71	13.95
	Vocabulary	78.08	11.60
Mid	Content	68.75	16.45
	Language	77.18	12.16
	Vocabulary	69.94	15.06
High	Content	67.13	19.10
	Language	81.90	12.13
	Vocabulary	83.86	15.06

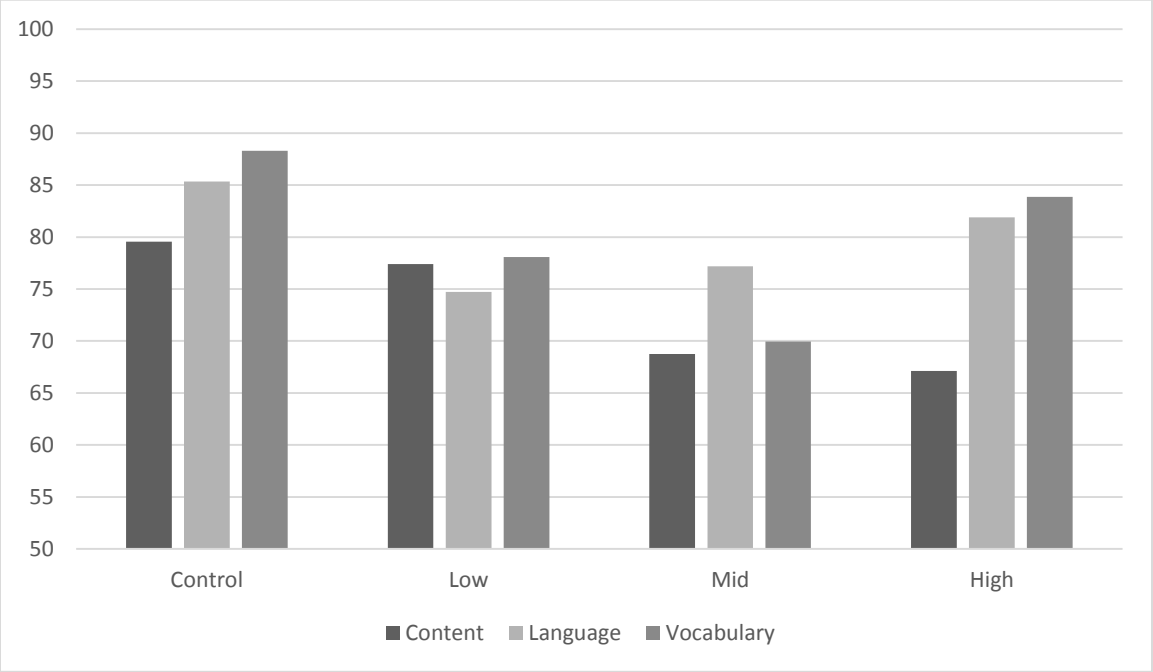


Figure 6. Comparing Mean Sub-Scale Scores with a Strict Error Interpretation

Post-Hoc Discriminant Function Analysis

In a MANOVA that utilizes independent observations, an additional post-hoc procedure, called discriminant function analysis can be used to look at the relationship between the dependent variables overall to determine if any of those relationships were also having an effect. In essence, a discriminant function analysis is a type of regression that uses the dependent variables to predict an essay's membership in the control, low, mid, or high experimental group. In discriminant function analysis, the dependent variables become the predictor variables and the independent variables are treated as a grouping variable.

The analysis revealed three discriminant functions, or combinations of sub-scale scales that act as a latent variable, significantly predicted an essay's placement in an experimental group. The first function explained 62% of the variance, canonical $R^2 = .29$. The second explained 32% of the variance, canonical $R^2 = .17$, whereas the third only explained 6% of the variance, canonical $R^2 = .04$. In combination, these three discriminant functions significantly differentiated between the amount of formulaic language in essays, $\Lambda = .57, x^2(15) = 55.08, p < .001$. By removing the first function, the second and third also significantly differentiated the experimental groups, $\Lambda = .79, x^2(8) = 22.32, p < .05$. When the first and second function were removed, the third function did not significantly differentiate the amount of formulaic language, $\Lambda = .96, x^2(3) = 3.99, p > .05$.

The correlations between sub-scale scores and discriminants (Table 32) revealed that vocabulary sub-scale scores loaded more highly on the first function (r=.84) than the second (r=.18) or third (r=.27). Organization loaded more highly on the second function (r=.38) than the first (-.40) or third (r=.23). Language loaded more highly on the third function (r=.75) than the

first (r=.44) or second (r= -.28). Finally, mechanics loaded more highly on the third (r= .36) than the first (r=.09) or second function (.08). These findings suggest that latent variable 1 distinguishes between the organization and vocabulary while latent variable 2 differentiates language from content.

Table 31.

Correlation Coefficients between Sub-Scale Scores and Discriminant Functions

	Function		
	1	2	3
Vocabulary	.84	.18	.27
Organization	-.04	.38	.23
Language	.44	-.28	.75
Content	.13	.59	.71
Mechanics	.09	.08	.36

Summary of Results

Even when a strict interpretation of comparable t-unit error rate was applied, manipulating the amount of formulaic language did not have a significant effect on the composite essay score (Figure 7). In an analysis of the sub-scale scores, formulaic language did affect the score, but not in a positive manner. Vocabulary showed the greatest number of differences between the levels; however, the control essay out-scored all the experimental conditions.

The discriminant function analysis revealed three underlying factors that affected the ESL Composition Profile (Jacobs et al., 1981) sub-scale scores, but only two of them were independently significant. Discriminant function analysis cannot explain which of the significant sub-scales- content, vocabulary, or language- were most responsible for the change in score. It

did, however, support the findings of Astika (1993) who found that vocabulary and content were the strongest predictors of score on the Profile. From the latent variables in the discriminant function analysis, vocabulary contributed the most to the first latent variable. Content contributed the most to the second latent variable's ability to predict group membership. Language contributed the most to the third discriminant function, but it is important to note that the third function was not a significant predictor of group membership on its own.

CHAPTER 5: CONCLUSION

This chapter discusses the results of the present study focusing on the significance and pedagogical implications of the findings. Before these key elements are presented, a brief review of the purpose, design, and results of the study are presented. The chapter also discusses limitations of the current study as well as implications for future research.

Purpose of the Study

At the post-secondary level, mastery of content area knowledge is most often assessed through writing assignments (Ferris, 2009; Knodt, 2006; Sullivan, 2006; White, 2007), which presents an added linguistic challenge for second language learners. These assignments tap into linguistic competence, but they also demand proficiency in academic language, which is a nuanced skill. Appropriate academic conventions vary between spoken and written scenarios (Biber, 1986; Biber et al., 2004; DeVito, 1967; Halliday, 1979; McCarthy, 1998) and between disciplines (Biber, 1986; Fang & Schleppegrell, 2010; Lee & Spratley, 2010). Even after extensive preparation in their home countries, and perhaps studying in Intensive English Programs, most NNSs continue to struggle to be successful writers in the university classroom (Ferris, 2009).

One area of interest to TESOL researchers is the role of vocabulary in second language writing. Vocabulary has been found to be an important predictor of second language writing success (Engber, 1995; Guo et al., 2013; Gonzalez, 2013; Grobe, 1981; Laufer & Nation, 1995; Santos, 1988). The influence of vocabulary on writing is seen across the student experience from standardized tests, to pre-admission programs, to the actual university classroom itself. For example, Guo et al. (2013) found that vocabulary accounted for 62% of variation in scores on

timed writing tasks for the TOEFL iBT. Engber (1995) found that IEP writing instructors were particularly sensitive to vocabulary-related errors, as lexical errors had a significant negative effect on writing score. Santos (1988) found that professors awarded harsher scores for content than linguistic conventions; however, in cases where vocabulary was the root of the error, the penalty on score was most severe. Second language learners are also attuned to high lexical stakes. They request more academic vocabulary instruction from their teachers during end of course surveys (Folse, 2010; Leki and Carson, 1994), ask for lexical guidance during in-class writing sessions (Gonzalez et al., 2012), and add and delete words based on their individual perceptions of what vocabulary will make their paper "more academic" (Coxhead, 2012).

In light of this role of vocabulary in second language writing, researchers, material writers, and instructors turn to corpora to decide which words in particular give learners the most return for their cognitive investment, especially in light of the pressure NNSs' face to show dramatic gains in proficiency even after receiving relatively small amounts of instruction (Cobb, 1999; Folse, 2010). To streamline this process, vocabulary lists such as the General Service List (West, 1953), the Academic Word List (2000), and the New Academic Vocabulary List (Gardner & Davies, 2013) identify items that have been deemed beneficial to learn. However, as more sensitive tools are developed to analyze corpora, researchers have discovered that the lexical hallmark of fluent academic language may not be held solely in single words like *insight*, which is a headword on the AWL, but perhaps in extended patterns of words such as *insight into the*, which is a formulaic sequence from the Academic Formulas List (Simpson-Vlach & Ellis, 2010). In other words, producing academic language may require particular words combined in particular ways.

While patterns of vocabulary words are common to all languages, the formulaic nature of the English lexis is estimated to account for between 30 and 50% of language production (Biber et al., 1999; Conkin & Schmitt, 2012; Howarth, 1998). Researchers suggest recurring patterns of language are so ubiquitous because they help speed up processing and production for native speakers (Arnon & Snider, 2010; Pawley & Syder, 1983; Schmitt et al., 2004). For non-native speakers, formulaic language has been termed "islands of reliability" (Dechert, 1984, p. 227) that help ease the burden of language production. Yet despite the benefits to both groups, empirical evidence continues to point towards a large gap in formulaic language production between native and non-native speaking counterparts (Granger, 1998; Howarth, 1998; Yorio, 1989) that is difficult to bridge even after instructional interventions (Jones & Haywood, 2004; Schmitt et al., 2004). Studies by Boers et al. (2006), Ohlrogge (2009), and Stengers et al. (2011) suggest that this gap is important, as formulaic language production significantly correlated with NS judges of written and spoken proficiency. To further complicate the matter, studies on the most salient formulaic patterns in academic language are not particularly pedagogically-minded as they produce thousands of sequences (Altenburg, 1998; Biber et al., 1999), some of which (*is in a, is one of the*) are difficult to envision in teachable moments.

The purpose of the current study was, therefore, to discover the effect of integrating formulaic sequences from a corpus-informed, pedagogically-oriented list on writing score. As NNSs walk a balance beam between the demands of producing correct and natural language, it is unclear how formulaic language fits into the discussion of effective L2 writing instruction, and ultimately the role it should play in preparing NNSs for success in higher education.

Summary of Findings

Four non-native speaking graduate students at a large research university in the United States generated academic writing samples. The writing samples were based on four argumentative prompts each assigned to an experimental condition related to the amount of formulaic language. The first prompt acted as the control. The second, third, and fourth prompt included a list of eight, 16, and 25 target formulaic sequences respectively taken from the Academic Formulas List (Simpson-Vlach & Ellis, 2010). The NNSs were asked to incorporate the target sequences into their response. The resulting essays, described in Table 10, were assembled into four testing packets. Each testing packet contained all four experimental conditions of formulaic language and incorporated one essay from each NNS (Table 9). These four core packets were then reordered into a total of 96 combinations, 24 possible combinations for each core packet, to control for order of grading effects.

The testing packets were randomly assigned to ESL writing instructors (*n=102)* for scoring during on-site data collection at eight IEPs in the southeastern United States (see Table 4). The ESL Composition Profile (Jacobs et al., 1981), or simply the Profile, has five sub-scales that measure content, organization, vocabulary, language, and mechanics for a composite score out of 100. These sub-scale and composite scores were the dependent variables. The first research question uncovered the effect of formulaic language on composite Profile score. The second looked at the scores on the individual sub-scales to determine if there was an effect for amount of formulaic language that might be overlooked by analyzing the composite score alone.

Research Question 1

A repeated measures ANOVA was used to determine whether there was a significant difference in Profile composite score between the control and three experimental amounts of formulaic language. The results showed a statistically significant difference in score, but the directionality of the change was unexpected. For the control and low conditions, the addition of 8 target sequences, or 5% of the words in the essay, did not cause a significant change in score. The addition of 16 and 25 sequences also did not significantly change the score between the mid and high conditions. There was, however, a significant decrease in composite score between the control/low and mid/high conditions with the control and low essays outperforming the others. An ANOVA was re-run on the composite scores of a smaller set of four essays selected based on the ratio of error-free t-units (Table 26). When the proportion of errors was more tightly controlled, the composite scores for each experimental level also decreased in relation to the control, yet the drop was not significant. In sum, the essays that included increasing amounts of formulaic language did not have a scoring advantage when graded by ESL writing teachers regardless of the amount of errors in the essay.

Research Question 2

Based on the results of a repeated-measures MANOVA, a significant difference was found between sub-scale scores. In the post hoc one-way ANOVAs, two scoring groups appeared with the control and low condition significantly outperforming the mid and high condition (Figure 3) in relation to the content, organization, and language sub-scales. For the vocabulary sub-scale specifically, there was more variation in score resulting in a slightly different pattern. While there remained no significant difference between the control and low

sub-scale scores, the control essays only outperformed the mid level essays. The low essays outperformed both the mid and high essays. In terms of mechanics, there was only one significant difference in sub-scale score between the low and mid level essays. The other experimental conditions, namely control, low, and high, were statistically similar.

Using the same four essays with a stricter control for proportion of error-free t-units from research question one, the presence and amount of formulaic language did not improve sub-scale scores across the conditions. Based on the results of a discriminant function analysis, the sub-scale scores were affected by three latent variables, only two of which were able to predict group membership in an experimental condition by themselves. The first latent variable had a strong correlation with vocabulary. The second latent variable had a strong correlation with content. In other words, vocabulary and content scores, not language or mechanics, seemed to be the strongest predictors of which experimental condition an essay belonged.

Significance of the Findings

Previous studies suggested that non-native speakers, even at the advanced level, use less formulaic language (Durrant & Schmitt, 2009; Granger, 1998; Hewings & Hewings, 2002; Li & Schmitt, 2009; O'Donnell, Römer, & Ellis, 2013) and tend to overuse a set group of formulaic sequences (Granger, 1998; Hewings &Hewings, 2002; Howarth, 1998) which serves to categorize their writing as novice by seasoned members of a discourse community, who in empirical studies are often native speakers (Altenberg, 1998; Cortes; 2004; Granger, 1998; Moon, 1998; Wray, 2002). In the current study, the number of formulaic sequences used in advanced non-native writing samples were manipulated in order to determine whether the

sequences could increase the score at both the composite and sub-scale level. The results indicated that this attempt was unsuccessful.

The findings of the present study directly contradict previous empirical results. In Ohlrogge (2009), intermediate-level NNSs' (*n=170)* scores on an in-house proficiency test significantly correlated with the amount of formulaic language used in their timed writing assessments. Yorio (1989) conducted a case study on one native Korean speaker and found that formulaic language played an important role in improving the comprehensibility of his L2 writing despite morphological and syntactical errors. While not studying written language production, Boers et al. (2006) and Stengers et al. (2011) found that formulaic language played a significant role in NS judges' perceptions of oral fluency. It is important, therefore, to consider why the present study produced such dramatically different results.

They key to this question is most likely found in a comparison of the operational definitions of formulaic language. Wray (2002) noted that there are over 50 terms and conceptualizations of formulaic language found in the field today. What the above mentioned studies have in common is that formulaic language was allowed to take on multiple forms. For example, Ohlrogge (2009) included eight categories of formulaicity varying in length and semantic transparency such as collocations, idioms, phrasal verbs, personal stance markers, transitions, copied language from the writing prompt, writers' biographical information, and general rhetorical structures. Yorio (1989) used collocations, idioms, and set phrases to describe formulaic language in the writing of his subject. Boers et al. (2006) and Stengers et al. (2011) used NS intuition to identify collocations, idioms, and sequences that approximated target-like production. Such widespread operational definitions of formulaic language make it difficult to

pin down what qualifies as "an appropriate range of multiword units" (Cowie, 1992, p. 11), as it is unclear if each category is equally responsible for perceived proficiency gains.

The current study, therefore, relied on a corpus-based list of academic formulaic language in order to generate the operational definition and select target sequences. Corpus-based lists of single-item words are based on the premise that not all words increase decoding power to the same degree, so large, purposely-collected samples of language can reveal important information about how a target-audience uses vocabulary (Coxhead, 2000; Nation, 2001a; West, 1953). Furthermore, the results of corpus-based analyses allow researchers, material writers, and teachers to present natural language in context without relying on human intuition, which can be unreliable (Folse, 2004; Nation, 2001a; Sinclair, 1991; Tsui, 2004). Finally, word lists inherently narrow down and prioritize vocabulary learning goals for language learners faced with the daunting task of acquiring a range of 7,000 to 10,000 words in order to function in the L2 (Hu & Nation, 2000; Nation, 2006; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, 2008). In the study of formulaic language, the generation of a pedagogically-minded list was especially important as watershed reports such as Biber et al. (1999) and Nattinger and DeCarrico (1992) produce valuable, but extensive taxonomies of sequences.

The target sequences used in the current study were taken from the writing sub-list of the Academic Formulas List, or AFL (Simpson-Vlach & Ellis, 2010). The AFL used teacher perceptions of formulaic language to triangulate traditional corpus tools such as frequency, range, and a measure of cohesiveness between words in a phrase. The AFL methodology focused specifically on generating a pedagogically useful list that would be appropriate for inclusion in a program of study that prepared academically-oriented NNSs (Simpson-Vlach &

Ellis, 2010). The sequences in this study (Table 33) were "formulaic, coherent, and perceptually salient" items (Simpson-Vlach & Ellis, 2010, p. 508) consisting of a uniform length and further identified as frequent in an independent corpus of university-level A-grade papers. The sequences are not idiomatic in nature, and are overwhelming made up of general vocabulary. In fact, only two of the words within a target sequence appear on the AWL: *insight, period*. The rest are considered high frequency, general vocabulary (Cobb, 2014).

When the operational definition of formulaic language was constricted to include only the bundles described above, the presence of formulaic language, regardless of the amount of sequences incorporated, did not give advanced L2 academic writers an advantage. Experienced ESL writing teachers did not score their papers as having better content, organization, vocabulary, or control of language conventions, nor did they receive higher scores overall. It is important to note, however, that these results do not counter the methodology used to generate the AFL or the validity of its findings. These sequences are certainly salient based on the multitude of criteria used to identify them. The current findings do, on the other hand, suggest that "no doubt certain lexical chunks need to be mastered for certain kinds of pragmatic competence; but we need to know which chunks, for what purposes" (Swan, 2012, p. 117). In other words, the quality of psycholinguistic salience alone may not identify formulaic sequences that are paramount for L2 learners to acquire and produce. Swan (2012) also suggests that NNSs may benefit from knowing formulaic sequences relevant to communication in specific situations or disciplines, but, as the current study suggests, a general purpose inventory of lexical patterns that appropriate native-like linguistic choices are not a prerequisite of fluent and effective communication.

Table 32.

Target Sequences Used in Low, Mid, and High Experimental Levels

Target Formulaic Sequences	
it is possible	it is necessary
depending on the	this does not
this means that	it has been
in some cases	to do so
should also be	if they are
be explained by	be noted that
allows us to	insight into the
same way as	it is interesting
it is important	to carry out
should not be	it follows that
does not have	over a period
is affected by	at this stage
in both cases	

Advanced Non-Native Writers and Prior Knowledge of the Target Sequences

With the exception of two cases, the target sequences did not appear in the essays unless they were explicitly assigned as part of an experimental condition. Participant A included two target sequences (*it is possible, should not be*) in the control writing sample and one in the low condition (*if they are*). Participant C included one target expression in the control (*it has been*) and used the same sequence (*be noted that*) in both the low and mid condition. The inclusion of these sequences did not adversely affect the operational definition of intensity of formulaic language. However, the lack of occurrence in any of the samples without explicit instruction suggests that these particular combinations of high frequency words may not be common in their writing outside of the research setting.

While the individual words are general, high frequency items, the way in which they combine to become formulaic may have gone unnoticed by these L2 learners because they did

not need to decode them as a unit, which is an integral step in building the formulaic lexicon (Ellis, 2008; Schmidt, 2001; Wray, 2002). According to the connectionist theory of SLA (Ellis, 1998; 1999; 2003; 2005), noticing is needed before the benefits of repeated exposure can integrate the sequence into the interlanguage. The acquisition of formulaic sequences is also unpredictable as it depends on the individual's decoding needs (Wray, 2002). Perhaps the high frequency nature of the individual words within the sequences masked the formulaicity for these advanced L2 learners. One may argue that if these sequences are not known to be formulaic to these non-native writers, and the current study did not include explicit instruction on the meaning and appropriate use of the sequences before the writing sessions, the results might actually reflect a lack of skill in using the sequences, not an inherent lack of benefit in the sequences. However, language scores, which incorporated errors, were not significant predictors in either of the two discriminant functions identified in the MANOVA post-hoc. Also, in looking at the intersection of target sequences and error-free t-units, the syntactic or semantic misuse of formulaic language was not widespread. In fact, t-units that contained a target sequence were almost twice as likely to be free from syntactic or semantic errors.

Target Formulaic Sequences and Error-Free T-Units

Wray (2002) noted that formulaic sequences, while stored and retrieved holistically, are still subjected to an individual's interlanguage system and consequently any potential production errors that might arise. Across all experimental conditions and all the non-native writers, there were 196 opportunities to use the target formulaic sequences. In looking at the writing samples, the writers overwhelming used one target sequence per t-unit. There were only three cases, as

shown below, where a writer combined multiple target expressions into one t-unit. Therefore, the total number of t-units that included a target formulaic sequence in the study was 193.

- *It is important to be noted that both are the effective approaches if they are appropriately applied (A4)*
- *To do so, a qualitative study is carried out over a long period of time (A4)*
- *This means that it is possible that in some cases, a person reacts differently to a situation which might not be the same way as the way another person reacts (C4)*
- *It follows that doing a qualitative research gives us an insight into the individual's world over a period of time*

The researcher looked at how many target bundles were used correctly across the 16 writing samples compared to the linguistic accuracy overall. On average, a little over half of all the t-units, or 57%, in each paper were free from syntactic and semantic errors (Table 10), regardless of the incorporation of a target bundle. In looking only at t-units that contained target sequences, the formulaic sequence only intersected with a syntactic or semantic error in 22 of 193, or 11%, of the cases.

From these 22 cases, six were syntactic errors connected to or inside the formulaic sequence itself. There was no identifiable pattern relating to which sequences contained syntactic errors, except five of the six were found in the high experimental condition. Participant B was the only writer that did not make a syntactic error when using a target sequence. Examples of syntactic errors made in the writing samples include:

- *It will be interesting to know that bread could be harmful for your health (A4)*

- *It is necessary to recognize individual differences which allows to look for specific traits and unique characteristics of each human (C4)*
- *To do so professors and scholars in charge of research courses in graduate colleges do not have to show their bias towards one or the other and allow students to choose freely between the quantitative and the qualitative traditions according to their needs and preferences (D4)*

Sixteen cases of semantic errors connected to target sequences were found. Of these, six were connected specifically with the formulaic sequence *same way as* (Table 34). This sequence appeared in all three experimental conditions, and it appeared in 12 t-units. This means it was incorrectly used 50% of the time. Also, each non-native writer misused this expression.

Many three-word bundles are actually part of longer four and five-word bundles (Biber et al., 1999; Cortes, 2004; Hyland, 2008), and *same way as* is part of the longer bundle *in the same way as.* After looking the sentences where this bundle was incorrectly used, four of the six occurrences did actually use the longer sequence. The awkwardness resulted from the placement and/or extended collocations. For example, a review of the concordance lines for *in the same way as* in the academic sub-corpus of the Corpus of Contemporary American English (COCA) (Davies, 2008) showed a preference for the sentence structure *(modal) + [passive verb] + in the same way as + [noun]*, seen in the examples below:

- The standards would be enforced *in the same way as* any other federal law
- They would be assessed *in the same way as* everyone else

- The patterns of climate change are understood *in the same way as* hurricane mapping charts
- SES was scored *in the same way as* the Dutch sample

The non-native writers in the current study tended to use *in the same way as* as an equivalent to *the same as*.

Table 33.

Errors Involving the Target Sequence *SAME WAY AS*

Author	Bundle in Context
A2	…Learning a foreign language not the same way as learning one's mother tongue.
A3	The colleague evaluation may have their shortcomings too even if it may not work as the same way as the students' evaluation.
B2	Adults can definitely learn better than children or in the same way as children learn the language
C2	The only skill that adults may not be able to achieve in the same way as children is pronunciation
C3	It should be noted that in the same way as student evaluations and teacher self-evaluations, observations can be immensely beneficial
D3	More research is needed to allow us to make conclusive decisions as regards the use of surveys in the same way as they have been used

The sequence with the second highest number of semantic errors was *depending on the* (Table 35). This sequence was also assigned to all three experimental conditions, and it was incorrectly used 3 times, or 25%. A search of the COCA showed that the expression is common at the beginning or end of the sentence followed by the word *circumstances* or *situation* or by another sequence such as *the kind of, the nature of, the number of,* and *the type of.* None of these

collocations were included as target sequences in the study, although they do appear on the original AFL. The most common use of *depending on the* by the non-native writers was *depending on the fact that,* which had no occurrences in the COCA academic sub-corpus.

Table 34.

Errors Involving the Target Sequence *DEPENDING ON THE*

Author	Bundle in Context
B2	I disagree with the idea that only children can learn a foreign language successfully depending on the fact that they are young and fresh
B4	You should not be including any interpretation albeit depending on the numbers
D3	Depending on the fact that there is no obligation to take surveys students might not feel very positively motivated towards the task

The Change in Errors among Essays with Eight, 16, and 25 Target Sequences

Another potential explanation for the decrease in scores at the mid and high amounts of formulaic language is the notion that the non-native writers were too distracted with the task of using the bundles, and as a result their writing became less accurate. Returning to the percent of error-free t-units (Table 36), it is possible to evaluate the syntactic and semantic accuracy of the writing samples across the individual writers. Participant A's most accurate essay was the high condition and the least accurate essay was the mid condition. Participant B, C, and D all made the least amount of errors in the low writing sample and the most errors in the control essay. Based on findings from both research questions, the essays with the low amount of formulaic language received the highest scores overall but were not significantly different from the control essays.

Table 35.

Percent of Error-Free T-Units by Non-Native Writer

	Participant A	Participant B	Participant C	Participant D
Control	37	39	36	60
Low	49	78	73	71
Mid	36	63	72	70
High	57	49	54	61

Limitations of the Study

As with all research, there are limitations that result from balancing authenticity and control of experimental conditions. There are three limitations inherent with the design of the study. First, timed writing assignments were used as instruments as opposed to collecting classroom artifacts. Using classroom-based writing assignments would more accurately reflect the breadth of formulaic knowledge particular NNSs might have. However, these artifacts would not have controlled for the type or amount of formulaic language and, therefore, would have conflicted with the purpose of the study, which was to understand the potential benefits of incorporating a corpus-based formulaic word list in second language writing. Also, using timed writing assignments helped the researcher control for essay length, which is significant predictor of essay score (Engber, 1995; Ferris, 1994; Guo et al., 1981). In addition, this type of writing scenario is congruent with high stakes assessments and on the job writing requirements (White, 2007).

As a result of choosing timed-writing assignments, prompts needed to be developed for the study. The second limitation of the study is that each experimental condition was represented by a different argumentative prompt generated by the researcher. The researcher used four prompts in order to draw upon the four areas of knowledge growth for graduate students in a

TESOL Ph.D. program, as explained in Chapter 3. The main rationale behind using four prompts was to make it more difficult for teachers to directly compare essays in their testing packet and subsequently reduce order of grading effects (Spear, 1997). However, prompt 3, or the mid condition, was consistently the lowest scoring essay. None of the characteristics reported for the writing samples would indicate that the mid level essays were sub-par. The fall may, in some part, reflect the graders' bias towards the essay topic. Prompt 3 asked writers to argue for or against the value of student perception of instruction surveys to evaluate teaching effectiveness.

Finally, the study utilized convenience sampling. Convenience sampling limits the statistical generalizability of the results to other populations (Fraenkel & Wallen, 2009). However, all participants did have to meet specific criteria in order to participate. The researcher also collected data from eight Intensive English Programs at institutions ranging from large research to small regional and even private universities to include as much diversity in the sample as possible. Finally, the size of the sample (*n=102)* allowed for results with over 95% power, as seen in Chapter 4 and supported by G*Power.

Recommendations for Future Research

The majority of research on formulaic language in second language writing is conducted with advanced non-native speaking participants, usually in university settings after the NNS has gained admission to a degree-seeking program (Hewings & Hewings, 2002; Howarth, 1998; Yorio, 1989). The current study also follows this pattern. Research that incorporates beginning and intermediate proficiency levels is needed. In the present study, the integration of target formulaic sequences into advanced non-native academic writing did not significantly improve

composite or sub-scale scores. Future research is needed to determine if formulaic sequences identified using the same operational definition would benefit compositions written by beginning and intermediate level students. The "zones of safety" (Boers et al., 2006, p. 247) offered by sequences may have a different effect on ESL teachers' perceptions of writing quality when the composition has characteristics of developing linguistic competency. Furthermore, L2 writing development for college preparedness is not exclusively done in IEPs and beginning composition courses. K-12 schools play an equal role for NNSs with various proficiency levels. Replicating the study to include high school ESL writing samples and secondary English teachers would be beneficial, especially in light of the Common Core State Standards with its emphasis on academic and disciplinary language.

Second, a study that combines the current operational definition of formulaic language and the design of Boers et al. (2006) adapted for writing would be beneficial. In Boers et al. (2006), a quasi-experimental design was used to supply one class of NNSs with direct, explicit instruction of formulaic language found in class readings. The other class discussed single item vocabulary words and grammar but not formulaic language in the readings. It is unclear in the present study if the four participants, and subsequently the writing scores, would have benefited from instruction and examples of the target sequences.

A third suggestion for future research includes a more detailed look at the interplay between errors and formulaic language. Errors did not play a large part in the current design. Errors were classified into two general categories- syntactic and semantic. The number and type of errors were not altered in order to generate representative writing samples. The discussion related to errors was confined to the proportion of error-free t-units, regardless of the types of

errors found within. While the results showed no advantage for formulaic language even when errors were strictly accounted for, a design that can discuss the interplay between lexical and syntactical errors in detail would be beneficial.

Earlier in this chapter, a discussion is presented that associates the different operational definitions of formulaic language as a possible explanation for the conflicting results of the current study with previous empirical studies (Boers et al., 2006; Ohlrogge, 2009; Stengers et al., 2011; Yorio, 1989). More research is needed that slowly adds depth and variety to the operational definition. Empirical research that compares the effect of semantically transparent and semantically opaque formulaic sequences on the perceived proficiency of non-native writers at various levels of L2 mastery is warranted.

Finally, even though the current study showed no improvement in writing score among essays with different amounts of formulaic language, this does not definitively make a case against the importance of these target sequences for L2 learners. Research that investigates the perceived benefits of the target sequences in reading, listening, and speaking are also needed.

Pedagogical Implications

Native speakers acquire their L1 vocabulary slowly over time with over 18 years of input; however, second language learners need to master thousands of items in a much shorter time span, and in EFL scenarios, with much less input (Cobb, 1999). The use of corpora to assist L2 teachers, learners, material writers, and publishers in making strategic decisions about language instruction is not new. The Teacher Word Book (Thorndike, 1921) used a corpus of 3 million words to generate a list of 10,000 must-know vocabulary words for students. Ogden's Basic English (1930) consisted of 850 words essential for everyday communication and 150 additional

words for scientists. In 1953, Michael West created the General Service List that includes the most frequent words in English. The University Word List (Xu & Nation, 1984) served as the precursor to the Academic Word List (Coxhead, 2000). What has changed the landscape of corpus-based vocabulary lists are three critical factors. There are a growing number of freely available online corpora, such as the Corpus of Contemporary American English (Davies, 2008). Those outside the academic community have more access to and more training with online corpus-based tools such as Lextutor (Cobb, 2014) that help plan and track L2 vocabulary growth. Finally, within the research community, technology has generated more and more sensitive search features that aid in generating more and more refined lists (Gardner & Davies, 2013; Simpson-Vlach & Ellis, 2010).

The result of these converging factors can be seen in the increased pressure for teachers to use corpus not only in their lesson plans but directly in the classroom. A review of the International Teaching English to Speakers of Other Languages 2014 Conference program showed 12 sessions devoted to corpus linguistics, seven of which were specifically geared towards practitioners and classroom applications. In many cases, the increased use of corpus has been beneficial. For example, Tsui (2004) followed questions posted on a discussion board for EFL teachers in Hong Kong run by education scholars at a large public university. The discussion board allowed teachers to post and respond to questions. Tsui found that teacher-posted questions generally fell into six categories, all of which could benefit from corpus-infused information. The six categories included: (1) deciphering when to use words that appear to be synonyms, (2) determining if a words are in fact interchangeable, (3) understanding prescriptive and descriptive grammar-in-use, (4) identifying collocations, (5) learning situational or

stylistically appropriate language choices, (6) formulating examples when direct translations to the L1 are not clear. However, Swan (2012) notes that a growing enthusiasm for corpus and its potential to teach native-like structures may cause the field to lose sight of the fact that "most non-native speakers must…settle for the acquisition of a…relatively restricted inventory of high-priority formulaic sequences, a correspondingly high proportion of non-formulaic grammatically generated material, and an imperfect mastery of collocational and selectional restrictions" (p. 118). With the words of Swan and Tsui in mind, the pedagogical applications of the current study take shape.

Formulaic language, in the context of the present study, was represented by semantically transparent combinations of high frequency words. These target sequences were sampled from a corpus-based list and given to advanced non-native writers to incorporate into 30-minute timed writing exercises. For the 102 ESL writing teachers that participated in the study, the addition of these sequences, regardless of the amount, did not improve their perceptions of writing quality as measured by scores. As a follow up, 16 essays were randomly selected from the Michigan Corpus of Upper Level Student Papers, MICUSP. The papers were all argumentative essays written by native-speakers in graduate school studying literature, education, linguistics, sociology, or psychology. Most importantly, each paper in the MICUSP must have received an 'A' by the course instructor in order to be submitted to the corpus. A search was conducted to see how many of the 25 target sequences occurred in these 16 randomly selected essays. None of the essays contained more than four target sequences, and the coverage rate of running words in the text by sequences was under 1% (Table 37). While these 16 comparison essays represent a very small sample of the papers in the corpus, it does suggest that the kinds of formulaic

sequences included in the Academic Formulas List (Simpson-Vlach & Ellis, 2010) may not be essential in successful L1 or L2 writing. Ultimately, the pedagogical implications of the current study suggest that caution is needed concerning the role that corpus-based formulaic language should play in the L2 classroom. The findings suggest that teaching the sequences on the Academic Formulas List (Simpson-Vlach & Ellis, 2010) would not be a beneficial addition to an English for Academic Purposes curriculum.

Table 36.

Occurrence of Target Formulaic Sequences in 16 MICUSP Essays

MICUSP Essay ID	Percent of Running Words in Text	Target Formulaic Sequences Used
EDU.G2.02.1	.24%	to do so
EDU.G3.03.1	.27%	it is necessary; allows one to; it is possible, in some cases
EDU.G3.03.3	.07%	it has been
ENG.G1.02.1	.23%	should be noted; allows one to, in the same way that
ENG.G1.03.1	.14%	in some cases; it is interesting; does not have; it is necessary
ENG.G1.04.1	.21%	it is important, it is important, it has been, does not have
ENG.G1.05.1	0	--
LIN.G1.02.1	.49%	it is possible, to do so, it is necessary; should not be;
LIN.G1.03.1	.21%	it follows that, does not have
HIS.G2.04.1	.06%	in some cases, to do so,
HIS.G1.01.1	.16%	should not be,
SOC.G1.01.1	.10%	It is possible,
SOC.G2.02.2	0	--
SOC.G3.09.2	.25%	should not be
PSY.G2.08.1	.53%	it is possible, it is possible, it is important, it is important, it is important, to carry out, it is necessary
PSY.G3.03.1	.09%	should not be

Conclusion

This study looked at the potential benefits of integrating target formulaic sequences from a pedagogically-minded, corpus-derived list into advanced non-native writing. The results indicated that formulaic language did not have a statistically significant positive effect on the composite or sub-scale scores that measured composition skills such as organization, language, and vocabulary. Further analyses suggested that formulaic language did not benefit writing samples even when strict control of grammatical and meaning-based errors was applied. However, the analyses in Chapter 4 showed that the amount of formulaic language did produce moderate effect sizes, and based on the discriminant function analysis, there appeared to be two latent variables that predicted the amount of formulaic language an essay contained. However, these latent variables were shown to be strongly correlated to vocabulary and content, so it is not evident that unaccounted for variables altered the effectiveness of the target formulaic sequences and influenced the results. In sum, the study did not make a clear case for the immediate integration of formulaic language in L2 writing classrooms, yet it does generate more questions about the different definitions of formulaic language and a need to investigate each in turn.

APPENDIX A:
IRB APPROVAL LETTER

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Exempt Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Alison M. Youngblood**

Date: **January 03, 2014**

Dear Researcher:

On 1/3/2014, the IRB approved the following activity as human participant research that is exempt from regulation:

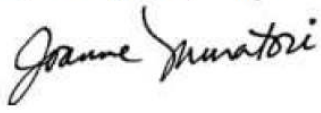
Type of Review:	Exempt Determination
Project Title:	Understanding the Effect of Formulaic Language on ESL Teachers' Perceptions of Writing Quality
Investigator:	Alison M Youngblood
IRB Number:	SBE-13-09918
Funding Agency:	
Grant Title:	
Research ID:	NA

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 01/03/2014 10:11:54 AM EST



IRB Coordinator

APPENDIX B:
NON-NATIVE WRITING SAMPLES

Essay A1 Prompt:

When making admission decisions, universities should stop using standardized tests like the GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

It is widely argued that when making admission decisions, universities should stop using standardized tests like GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work. Personally, I do not agree with the statement because standard tests will not affect the validity of admission and they do have some advantages in selecting candidates.

First of all, standard test could prevent corruption and cheating in some extent. The applicants may make false documents and portfolio. For example, they may ask others to write some writing samples for them, or bribe someone to make an appealing transcript. However, it is not very easy to cheat in a standard test. They will possibly be caught if they ask someone else to take the exams. Therefore, if the scores of standard tests are so different from what they appear to be, it will arouse people's attention.

Certainly, it is possible that sometimes people are feeling sick during the exams and do not perform their best. However, it will not affect the admission of real qualified candidates. I believe discussion among admission committee and some investigation will clarify the situation. The scores should not be a threshold to exclude qualified candidates but a good reference to find the best ones. Generally speaking, if a candidate is really as good as they describe, he/she will also do well in the standard tests.

Finally, the process of preparing for some standard exams sometimes is a very efficient way to enhance the academic performance of candidates in a short time. Take myself for example, my English reading speed was enhanced from 1000 words per hour to 10000 words per hour by preparing for GRE exams. For those who are really want to but make a hush decision to apply for graduate school, the GPA cannot be changed in a day; publication will not come out

that fast when people decide to go to graduate school. However, by preparing for standard tests, one will read more, write more and show his/her ability in studying and dealing with pressure, which I believe, is also an important quality for gradated study.

To sum up, the standard tests like GRE will not eliminate qualified graduate school candidates but help to admit qualified one. Moreover, it will enhance people's performance and prepare them for the future study. Therefore, I think it will still be a good idea to keep standard tests in graduate school admission.

End of Essay

Essay A2 Prompt:

It is too difficult to learn a foreign language when you are an adult. The only people who successfully learn another language begin studying as children.

Agree or disagree with the statement and provide support for your answer. You do not need

Many people think that it is too difficult to learn a foreign language when you are an adult and the only people who successfully learn another language begin studying as children. I personally do not agree such claim and I believe it is also possible for adults to be successful language learners.

It is true that in daily life, we may find more successful foreign language learners who start to learn foreign language as children than who start to learn them as adults. However, it can be explained by longer time devotion to language exposure rather than just age. The adults in most cases have to do job, housework and social activities, while children could be really devoted into language study when they want to. Therefore, it is possible for adults to achieve the same language proficiency or even better than children as long as they have equal time and language exposure.

Adults may have advantages in learning foreign languages over children. We all have memory that how easy when we learn our native tongue. However, without a natural language environment, learning foreign language is actually not the same way as learning one's mother tongue. Adults are more mature in cognition and logical thinking. This means that they are better at finding the rules and regulations in foreign language than children, and that is usually a really efficient way to learn language.

Another advantage for adults to learn a foreign language is that adults are more self-disciplined. Learning a foreign language is never an easy job. In some cases, it could even be tough and frustrating. However, strong will and self-discipline will allow us to persist and continue even when the learning process becomes really difficult. Children, on the other hand, might easily quit if they are not under the pressure of their parents.

Furthermore, one can never ignore the role of motivation in language learning. Carl Max started to learn Russia when he was fifty years old just because he was so enthusiastic to read those original Russian documents, and he was fluent in reading after six months. When people are really motivated, they can be successful in many things. Therefore, to be a successful language learner may also greatly depend on the motivation and how much one wants to be a successful language learner, not just age.

All in all, I guess it is safe to conclude that adults should also be able to become successful language learners, especially in the foreign language study, because they have more knowledge, they are more mature and they are more self-disciplined. Given a natural language environment, maybe children are better in catching precise pronunciations and intonation because they have not been so influenced by their native tongue. However, while studying a language from audio and textbook, adults will definitely have their advantage. Therefore, everyone should be confident to pick up a language at any age, and know they are very possible to learn it well.

End of Essay

Essay A3 Prompt:

The most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester.

Agree or disagree with the statement and provide support for your answer. You do not need

It is necessary to evaluate a teacher's performance regularly and there are different ways to convey the process. Some people hold the opinion that the most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester. I may not quite agree with that statement.

First of all, are students the most reliable group to evaluate a teacher? There is no doubt that it is important to include students' opinion in teaching evaluation, and there is no question that no matter how young a student is, they are able to tell a good teacher from an unqualified teacher. However, this does not mean that students are the most reliable groups. Students may be easily affected by their personal bias and emotion because they are very young. For example, a student may not give good scores to an immigrated teacher because his parents tell him the immigrants steal their jobs, or another student give good scores just because this teacher shares the same crush on Justin Bieber like her. In both cases, though they are able to know whether a teacher is good or not, they may not mature enough to be objective.

Secondly, is the end of the semester a good time for students to evaluate their teacher and thus influence the reliability of the result? It is possible that at this stage when under the pressure of final exam, students may be even less objective in evaluating their teacher. For example, a teacher getting a low score can be explained by not giving hint or good score for the students in final exams. In some cases, if a school relies too much on students' evaluation, teachers may design the course and homework more depending on what the students like rather than what they really need. This means that the evaluation does not have positive influence to teachers' performance but becomes a negative factor in the learning effect of the students.

Students' evaluation at the end of semester is definitely not the most reliable way for teachers' performance. Actually hardly any of the existed evaluation could be claimed as a most reliable one. For example, the colleague evaluation may have their shortcomings too even if it may not work as the same ways as the students' evaluation. Even one may find a most reliable evaluation, it may not be enough. In my opinion, to maximize the reliability of teaching evaluation, one should combine the results of evaluation from different accesses. The students' evaluation should not be the only way to evaluate teacher while occasional observation and peer review should also be included. Though none of these evaluations might be perfect but together they will allow us to provide a wholesome picture of how does a teacher do. After all, the aim of a valid teaching evaluation is to enhance teachers' performance and the quality of our education and not just about a class of happy students.

End of Essay

Essay A4 Prompt:

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provides hard evidence.

There are two approaches to conducting research—quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provided hard evidence, but I cannot agree with such statement.

First of all, both approaches are ways to do research. A tool should not be judged as superior or inferior by itself, just like one cannot argue fork is better than spoon without context. It is important to be noted that both are the effective approaches if they are appropriately applied. If a result is better explained by numbers, then quantitative research should be chosen and if it is done better by description, then qualitative research should be applied. The effectiveness of the approaches depends on how they are applied. In both cases, the approaches will allow us to explain the phenomenon or conclude the result of our research clearly to the others.

It is true that the document of qualitative study might be more affected by the researcher's bias or experience. However, this does not make numbers look more objective. It is possible that quantitative data is manipulated in the same way as the qualitative data. Also, it has been proved that pure statistical results in some cases can be really ridiculous. For example, it will be interesting to know that statistically bread could be harmful for your health and water is a fatal matter.

Another reason that qualitative approach should also be valuated is it does provide many details for the research. Quantitative data might be impressive in presenting result, but qualitative

data sometimes is really powerful at the stage of explaining reasons. It gives the advantage that simple number does not have. To do so, a qualitative study is carried out over a long period time and many important details are recorded on a single case. By doing so it provided us a deep understanding and insight into a certain phenomenon. It follows that sometimes qualitative is indispensable for the research to know better about what they are studying.

To sum up, it is necessary to be aware that both approaches are valuable for getting or explaining our research results. This means that when we are thinking which one to choose for our study, we should not have prejudice that if we do a qualitative study, our research will be underestimated. It is the match of research approach and research questions that does matter. We should not have superstition about numbers because numbers are not all objective and sometimes without the support of qualitative documents, numbers can mean anything or nothing. Moreover, qualitative data has its very advantage by providing thoughts, opinions, natural conversation and detailed description. Therefore, it is not an sound argument to say quantitative study is superior to qualitative one and it is not safe to say that numbers will provide hard evidence after all.

End of Essay

Essay B1 Prompt:

When making admission decisions, universities should stop using standardized tests like the GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

College students' ultimate goal during their undergraduate studies is sometimes finding a good job and sometimes getting in a graduate school. Regarding the graduate school, it is a very vital issue for them because preparing for standardized testing such as GRE and/or TOEFL is the biggest issue. However, what if a student only prepares for these types of standardized exams and does not know anything applicable in his/her field? What if this individual can do great on multiple choice tests and write perfect essays but bot speak even a little bit of English? There are two main reasons clarifying why graduate schools stop using standardized tests like GRE or TOEFL and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

First, standardized tests do not measure social skills and how individuals collaborate. Taking part in some assignments and duties in graduate school requires some collaborating skills. However, individuals taking the standardized tests do not collaborate with computer because it is a computer that they talk to but there is no response coming from the computer. Even if they ask something or want to assert an idea just to collaborate and solve a problem, they will not be getting any idea or any response from the computer. The individual is all-alone and cannot collaborate or solve a problem. Considering the fact that they practice and prepare for these exams for a few years, they will never and ever collaborate. So, what is the result? The graduate school will accept some robots that can only answer multiple-choice questions. However, if they present a portfolio based on what they do and how they solve problems or how they interact with people to create diversity, the graduate schools would make a good decision when they accept the best-portfolio-person in terms of accepting future's scientists or teachers who know the reality and find

solutions by interacting, collaborating and producing solutions for today's world.

Next, as it is known, such standardized tests do not measure the correctness of pronunciation appropriately. What if a person can only speak exam-type of language but not an interactive type of language? Linking this issue to the previously discussed issue, they would need to interact so that they can show how active they were all through their lives and prove the good daily life interaction skills. However, you can interact with a computer to a certain extent, which is you can state your idea and not get any idea from the computer. What if a professor does not understand you because of your pronunciation of words or accent? If the individuals do not study the pronunciation and intelligibility of a language that they are supposed to use during their graduate studies, then they can only succeed well in the exam but not in graduate schools. Instead, when a person submits a portfolio, graduate schools know what that individual attended and how can he/she speak a language.

End of Essay

Essay B2 Prompt:

It is too difficult to learn a foreign language when you are an adult. The only people who successfully learn another language begin studying as children.

Agree or disagree with the statement and provide support for your answer. You do not need

It is generally thought that it is difficult to learn a foreign language when you are an adult and the only people who can successfully learn another language begin studying as children; however, even though it is possible to learn a foreign language successfully as a child, adults sometimes can learn a foreign language better. Therefore, I disagree with the idea that only children can learn a foreign language successfully depending on the fact that they are young and fresh.

To begin with, even though children are young and it is easy for them to acquire a language, adults have better strategies and skills to learn the same language. For example, in one of the studies, the language-learning adult was compared to young children learning the same language and the adult's language results at the end of a three month period were better than the children. There might be some other effects in this study; however, even one study is enough to refute the idea above.

Second of all, children might not be motivated to study a language in a classroom but adults might be. This allows us to revise the idea of how young learners can learn a language easily. This can be explained by "motivation" to learn a language. If an adult is motivated enough to learn a language, he/she will do everything to learn that language. In one of the studies, Ms. Kaplan wanted to learn French and she was very motivated to learn it. She went to France for a short period of time and had a lot of friends speaking French to interact because she internally wanted to learn it. She was very successful and at the end of the study, it is stated that her French knowledge is very close to nativeness. This means that if an individual has enough motivation, he/she can learn a language very successfully regardless of their age. Therefore,

individual differences such as motivation should also be considered as one of the reasons why adults could also learn a language easily.

The third reason why adults can learn a language easily is that adults have the taxonomy of content knowledge and more world knowledge than young learners. In some cases, this knowledge helps adults to learn more about the culture of that language and helps their acquisition of that language. Considering the fact that language cannot be separated from its culture, one needs to learn about the culture of that language. Sociocultural factors are so important that they have vital effects even on learner variables including the ones I mentioned above. They affect the way learners approach a language as a target language.

Therefore, if adults' language learning strategies, skills and motivation are considered, and they are encouraged for being open minded towards the target language community by including sociocultural elements in their classes, adults can definitely learn better than children or in the same way as children learn the language.

End of Essay

Essay B3 Prompt:

The most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester.

Agree or disagree with the statement and provide support for your answer. You do not need

Being a teacher is a hard job if a teacher is doing his/her job properly. It is possible to see the hard job through some ways but the most reliable way is the students' feedback on the teacher's work. One cannot understand the quality of work done by observing that teacher for a short period of time or by looking at the classroom work of students. This means that if a teacher does not have the competence to be a teacher, understanding the quality of teacher can only be achieved through student surveys at the end of the semesters.

First of all, students' comments at the end of the semester allow us to learn more about the teachers' materials presented to students and their way of presenting them depending on the survey-questions. At this stage, the surveys can be really helpful. In addition, how much they know the content knowledge and whether they hesitate about some issues while explaining can definitely be understood through surveys. However, this does not mean that each survey is a very good survey. The questions in the surveys must be indicating the competence issues.

Second, teachers' attitudes can be explained by students' comments in the open-ended parts of the surveys. In some cases, it is not possible to understand teachers' attitudes from the multiple choice parts of the surveys; however, when given a chance, students "anonymously" give good feedback on what the teacher was doing and how they liked this behavior. However, as emphasized, it is important that the surveys should not be with students' names. In both cases (open-ended and multiple-choice), the surveys should not include students' names. When the surveys want names, students do not want to show reality or state the negative comments in case they might take a class from the same teacher again and if the teacher knows this student's name, then he/she might get a lower grade. Then, the result is affected by this factor. Therefore, it is vital for surveys to be anonymous and it is the best way to learn about teachers' competence.

Last of all, one can learn about teachers' responsiveness and punctuality in the same way as learning about the materials. Some scholars assert that this factor should also be considered under the idea of teachers' competence and some do not share the same idea. However, in my opinion, a teacher's responsiveness to students through emails or face to face interaction as well as punctuality of coming to the class in time or ending the class in time must be considered as a competence. It is necessary for a teacher to know the ground rules and comprehend them. You can also get this information through surveys at the end of the semesters.

To summarize, if a professional who is interested in learning more about teachers' competence should use the surveys that students fill in at the end of the semesters. They help to understand the material, teaching style, teacher attitude and responsiveness and punctuality of teachers.

End of Essay

Essay B4 Prompt:

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provides hard evidence.

There are two approaches to conducting research such as quantitative and qualitative. The first one uses numbers to measure effects and explain the results while the latter uses descriptive tools such as interviews. It has been argued that some people think quantitative research is superior to qualitative one because it allows us to present hard evidence, and I agree with this idea and this issue can be explained by two perspectives.

First of all, it is important to provide the reliability of a research and it is necessary to know that qualitative research cannot be reliable. It is possible to record the data and interpret it; however, how can a data that is interpreted be reliable or as reliable as quantitative method? This means that if you want your research to be the most reliable one, then you should try to find a way to apply the same type of idea as a quantitative research and should also be giving an insight into the topic that you are searching. It should be noted that, in some cases, it is not possible to have both qualitative type of research and reliability. Then, why are you trying hard to apply a qualitative way? For providing reliability to the audience, you need a quantitative way, which consists of numbers. Obviously, your reliability of data is affected by interpretations, so you should not be including any interpretation albeit depending on the numbers.

Second of all, qualitative method might be the best in terms of validity; however, this does not mean that quantitative research data is not valid. Quantitative data can be valid in the same way as qualitative one if you provide appropriate questions in a questionnaire. To do so, one should consider doing a pilot research to find the right questions and interview with the participants of the pilot study after they answer the questions. If they are not consistent about

some answers, you should do some revisions because the answers might change over a period of time. Even though quantitative research requires good preparation, qualitative one does not have this type of a preparation period. This also shows us why researchers should choose quantitative research methods. Some might say that quantitative one needs a lot time and it is a waste of time. However, at this stage, the quality of questions needs to be considered. In qualitative research, it is very interesting for someone to talk about the quality of his or her research. How can you prove the quality of your research by asking only one question or a few questions? To carry out a good research, you need to ask a lot of questions and this can only be done through quantitative method.

To summarize, there is no point in doing a qualitative research if one needs to be both valid and reliable because it follows that researcher needs to apply a rule to provide validity and reliability in both cases, but these can only be achieved through quantitative research.

End of Essay

Essay C1 Prompt:

When making admission decisions, universities should stop using standardized tests like the GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

Getting admissions from universities is getting more and more difficult every year as the number of applicants increases and more jobs require university degrees and academic skills. Naturally, as the number of applicants goes up and there is a big pool to choose from, universities need to come up with stricter gatekeeping and admission requirements to be able to narrow down the candidates. Standardized tests such as the GRE are one of the admission requirements that most universities request in order for the applicants' documents to be reviewed for admission. However, I believe that using a portfolio of scholarly and professional work would be a better alternative to submitting GRE scores for the following reasons: the GRE exam only examines test-taking skills, it is not a good representative of an applicant's academic writing skills, and also by seeing the applicant's actual academic work universities and faculty can have a better understanding of whether the applicant can be a good fit for their program or not.

It has been argued throughout the years that the GRE exam only tests test-taking skills. Every year many test-takers spend a lot of money on preparation classes and materials for the GREs. Having seen many people taking the GRE preparation classes and going through many advertisements and information sessions for such classes myself, I know that one of the topics discussed in most of these classes is about tricks and how to find the right answer with different test-taking skills to eliminate wrong answers, look for wording clues in the questions etc. Therefore, the majority of time spent for preparation for the GREs is spent on test-taking skills rather than knowing the materials.

Moreover, the GRE scores are rarely a representative of a student's actual academic skills. The test consists of three parts: analytical, verbal and quantitative. Without the

consideration of what major the test-taker is specialized in, he or she has to take all the sections of the test and it will affect his or her overall score. Unfortunately, some schools require a minimum overall GRE score for applications to be reviewed and pass the first step of admission and this issue can affect many applicants negatively. For instance, if a person has been majoring in the Humanities, he or she might not have had any math classes in years and will probably never need to know math in his or her career, as a consequence he or she might not do well on the test overall and might not be considered for admission only because his or her overall score did not match the requirements.

Furthermore, seeing an applicant's portfolio of scholarly and professional work can give the reviewers a better insight on the applicant's future in the program. The writing section of the GRE is not about the topics that most students would write about in their profession and major and therefore is not a good depiction of their success in the field.

Overall, submitting a portfolio is a better alternative to taking the GRE exam because it only tests test-taking skills, is not a good example of the test-takers academic writing or abilities, and cannot predict their future in their studies and how successful they will be in the programs they are applying to.

End of Essay

Essay C2 Prompt:

It is too difficult to learn a foreign language when you are an adult. The only people who successfully learn another language begin studying as children.

Agree or disagree with the statement and provide support for your answer. You do not need

One of the main questions issues teachers of English as a second or foreign language are faced with is that students, or even in some cases teachers themselves, believe that it is too difficult to learn a foreign language when you are an adult. They believe that children can only learn the language successfully and achieve native-like proficiency. However, I disagree with this statement for three different reasons: the only part about learning a foreign language that might be easier for children is pronunciation, children have a much more limited need for vocabulary and complex sentence structures, and also adults know their learning techniques better and have better control over their progress and difficulties.

First, most adults argue that they can never speak a foreign language in the same way as children do. However, speaking a language consists of many different factors. This means that language is not limited to everyday social conversations, or writing an academic paper, there are many factors that come into play that allow us to divide languages and learning skills. Generally, learning and teaching a foreign language is based on four different skills: speaking, listening, reading, and writing. Biologically speaking, research has shown that depending on the learner, the only skill that adults may not be able to achieve in the same way as children is pronunciation. This can be explained by the movements of the tongue muscles, as we grow older our tongue muscles get used to certain movements that our first language requires and it can make it more difficult to adults to adapt to the pronunciation of some sounds in the second language which are different from their first language.

Moreover, the way children use language is different from the way adults use it. When we look more closely, children's exposure and usage of language is limited to very short, simple conversations or even chunks at times without the need to have complete grammatically correct

sentences. On the other hand, adults' language use is much more complex and on a daily basis an adult can have conversations or written correspondence about various topics such as their academic subjects, politics, sports, work, family etc. Therefore, the spectrum of language usage is much broader for adults than children and consequently they cannot be compared in the same way.

In addition, it should also be noted that average adults have had years of schooling and training which makes them more knowledgeable and aware about their learning habits. Thus, adults can use their learning skills in order to break down language structures and learn them in an easier way. However, this might take years of practice for children to achieve.

Overall, even though it is possible that an individual learns a language better than another individual, age is not the main factor in learning a foreign language. Children's limited usage of language, adults' knowledge of their learning skills, and biological evidence show that both adults and children can learn a foreign language.

End of Essay

Essay C3 Prompt:

The most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester.

Agree or disagree with the statement and provide support for your answer. You do not need

Teacher evaluation is one of the issues that many institutions and universities have debates on. Some argue that the most reliable measure of a teacher's competence is the anonymous student surveys that the students fill out at the end of the semester. However, this does not mean that student surveys are the only or most reliable measure. There are other ways that we can check teacher's teaching competence which can be more valid and a combination of these evaluations can benefit educators more in the long run, such as teacher self-evaluations, observations, and providing good management and working environment.

It is possible that student surveys give us a good overview on how students perceived the class and the teacher's teaching competence, however, it is important that we take teachers' self-evaluation into account as well. Depending on the types of courses, time of class and the number of students, a teacher can choose to pursue different teaching strategies. For example, in some cases the teacher might choose to be stricter with one class or give more or less assignments to the next class. This means that considering teachers' self-evaluations allows us to have a better understanding of teaching and also puts teachers in check with how they teach in class and handle different situations.

It should also be noted that in the same way as student evaluations and teacher self-evaluations, observations can be immensely beneficial to teachers' competence. Students may sometimes make comments that are not necessarily profession or based on the teacher's teaching abilities. Also, students' opinion is sometimes affected by the pressures of the end of the semester. Thus, it is necessary to have experts' opinions as well. These observations can be done by other teachers who can evaluate the teaching competence of the teacher with an open and experienced eye. The observation does not have to be very minute or it should not be only focused on the

negative or not very good strategies the teacher used but it can be an overview of both strong and weak points. In both cases, observations can evaluate and improve an educator's future performance.

Moreover, management plays an important role in teachers' competence and evaluation. Sometimes, student evaluations of a class might be based on issues that are not under a teacher's control. For instance, if there is a strict curriculum that a teacher has to follow that does not benefit the students much then students might not give good feedback on how the class was planned and the lessons that they had to cover. This can be explained by lack of good management and supervision. At this stage, managers and supervisors monitor the content of courses and find curricula in order to achieve better results and student satisfaction.

Therefore, student surveys can be used as one way of evaluation but are not the most reliable source to measure an educator's competence. Teacher self-evaluations, observations, and good management can also be beneficial.

End of Essay

Essay C4 Prompt:

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provides hard evidence.

The two approaches to conduct research, quantitative and qualitative, have always had their supporters or opposers for each method. Even though in both cases there are advantages and disadvantages depending on the topic, I lean more towards the qualitative method in the humanities majors for two different reasons: the nature of the subjects of study, and limitations of quantitative research to generalize the findings.

It is important to know that the subjects of a study in majors related to humanities are people and as humans, we are each different and unique. This means that it is possible that in some cases, a person reacts differently to a situation which might not be the same way as the way another person reacts. It is necessary to recognize individual differences which allows us to look for specific traits and unique characteristics of each human. This does not mean that we ignore the fact that humans are social beings and live as a group and whole, but at the same time it should not be considered that everyone responds the same way as everyone else. It follows that doing a qualitative research gives us an insight into the individual's world over a period of time and to do so gives value to humans both as an individual and as a whole.

Furthermore, it has been argued that quantitative research is better to carry out because it provides hard evidence and uses numbers to measure effects. It is interesting to see both sides of the issue, however, there are some limitations to conducting quantitative research. Overgeneralization is an issue that quantitative researchers and the general audience of research struggle with; findings of a research cannot be overgeneralized to a big group of people. For instance, there was a research on the effect of taking an aspirin daily and how it avoids heart

attacks. The problem with that research was that the subjects were white young medical student men which are not a good representative of every race and people around the world in general. However, this research's findings were all over the news and we might still hear people saying that our health is affected by medications and it is good take an aspirin daily to prevent heart attacks. Therefore, at this stage it is important to be noted that the research design is very crucial and the findings of research on one group may not be true for another group and the same phenomena cannot be explained by the same methods.

Overall, both quantitative and qualitative research have their own strength and weakness and it is up to the researcher to see if they are beneficial to his or her research question. Qualitative research pays more attention to the individuality of humans and it does not have to look at humans as a group and it should also be taken into account that qualitative method mainly does not overgeneralize the findings to a much larger group of people.

End of Essay

Essay D1 Prompt:

When making admission decisions, universities should stop using standardized tests like the GRE and instead ask graduate school applicants to submit a portfolio of scholarly and professional work.

Graduate school applicants in the United States of America have to undergo a stressful time when applying to continue their studies in graduate school. Apart from filling in extensive application forms and submitting a great amount of documents, which include previous scholarly work completed in a previous academic stage, they should prepare, practice and sit for a very demanding standardized test called Graduate Records Examination (GRE). This requirement adds up to the already difficult and complex set of requirements plus increases students' anxiety levels.

Proposals against this laborious admission process have been released in the last years. There are certainly different alternatives that can be taken into account at the time of evaluating future graduate students. First, it should be taken into account that when applicants decide to continue their education in graduate school, they have already finished previous stages in their education and have therefore produced scholarly work that can be used as a means of evaluation of their academic abilities to attend graduate courses and accomplish academic work successfully. Second, the work they have already produced in the past has been re-read, revised and edited, consequently it reflects the best ability the candidate possess to produce work that has been graded against academic standards. Third, previous work should not only reflect abilities the candidate may have in one skill, but it can also show abilities in different areas of expertise.

Consequently, I fully agree with the proposal explained above because as a graduate student who has had to undergo many tests and obstacles to attain my goal to attend graduate school and obtain my PhD degree, I would feel much more respected and better evaluated if my previous achievements were taken into account as well as standardized tests. Even when performing well in these tests is a proof of academic ability and higher order thinking, I still

think they cannot really measure a persons' academic ability or commitment to the tasks he or she has set out to accomplish. There is no test that can offer a real picture of a candidate's abilities and knowledge and most importantly, there is no test that can do this without causing great amounts of stress, which does not add up positive elements to the final image evaluators can form of the applicant.

End of Essay

Essay D2 Prompt:

It is too difficult to learn a foreign language when you are an adult. The only people who successfully learn another language begin studying as children.

Agree or disagree with the statement and provide support for your answer. You do not need

Research in Second Language Learning (SLL) and Second Language Acquisition (SLA) has demonstrated varied results as regards the ability that adults possess to learn a Foreign Language (FL) or a Second Language (SL). As it is explained by researchers and practitioners, adults may have different experiences in this process and this means that results can be different. It is possible to find cases of adults who have been so successful in their FL learning that they can be judged to be native speakers by other native speakers of that language. On the other hand, there are cases of learners who can barely be understood in their interlanguage, who really do not make much progress in their learning process and whose final results in this process cannot be considered to be successful.

Researchers have found out that depending on the environment in which these learners live, practice and learn on a daily basis their results in learning a FL or a SL will be more or less successful. In some cases we know that FL or SL learners are supported by their environments to learn and this can be considered to be a positive influence in this process. Take for example the case of adults who should work in companies or offices in which the FL or SL is the main means of communication among colleagues, or between bosses and employees. These kinds of environments should also be very conducive towards positive learning results, because the individual is constantly supported by an environment in which he or she needs to use the language in order to communicate successfully. Studies of these kinds of milieus allow us to focus on the need individuals experience to learn the FL or SL, but at the same time, they offer us the possibility to discover cases in which these same environments can cause stress on the individual, putting much pressure on him or her to learn the SL or FL. In this last cases, results

cannot be considered to be so successful due to the fact that pressure adds a negative element to the experience.

Some studies have also been carried out in environments which are not so conducive to FL or SL learning in the same way as they have been conducted in the positively conducive environments. In the less positive environments, adults are found to be slower in their learning process because they do not have the need to learn and use the FL or SL as the main means of communication. This occurs in communities where speakers of the same native language meet or live together for extensive periods of time and therefore use mostly their native languages for everyday conversation. In these cases, individuals may attend FL or SL courses but only practice these in the classroom environment for learning purposes. These learners' experiences can be considered to be positive, especially in cases in which they are not pressed to use the FL or SL to accomplish job duties, but their results in the learning process cannot be considered to be very positive.

End of Essay

Essay D3 Prompt:

The most reliable measure of a teacher's competence as an educator is through anonymous student surveys given at the end of the semester.

Agree or disagree with the statement and provide support for your answer. You do not need

Universities in the United States of America are used to administer anonymous student surveys at the end of each semester so as to evaluate the professors' competence as educators. As it is explained by research conducted in the area, it is important to consider that students are not obliged to complete these surveys so it is possible that some professors get responses from all of their students, some others only from some students, and still others who do not receive any feed back. Therefore the analysis of the results obtained is affected by the fact mentioned above, i.e. that survey taking is not compulsory.

Researchers and practitioners have established some research questions as regards the importance that should be attributed to the use of surveys as instruments for evaluation. This does not mean that surveys and their results should be disregarded, but their analysis should be carried out according to certain constraints. Research is not conclusive at this this stage as regards whether surveys should be disregarded altogether or whether they should be taken more seriously. It is necessary to consider different factors in both cases, when administering surveys and when analyzing surveys' data.

On the one hand, the not compulsory nature of survey taking should be considered one of the top elements to be taken into account in this discussion. Depending on the fact that there is no obligation to take surveys students might not feel very positively motivated towards the task. In some cases, they might even take them carelessly, not paying the needed attention to the information they are providing. On the other hand, it should also be taken into account the fact that anonymity is not always a positive feature because some survey takers become more careless and do not always express their real opinions and ideas by these means. This means that, surveys do not always provide information we can surely rely on.

Results of investigations should not be considered as conclusive results of the way instruction is imparted at universities. More research is needed to allow us to make conclusive decisions as regards the use of surveys in the same way as they have been used, or in different ways. In conclusion we do not have to be driven by the results that have been produced so far and continue investigating this issue.

End of Essay

Essay D4 Prompt:

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and journals to answer a question. Some people think quantitative research is superior to qualitative research because it provides hard evidence.

There are two approaches to conducting research- quantitative and qualitative. Quantitative research uses numbers to measure effects. Qualitative research uses descriptive tools such as interviews and qualitative research because it provides hard evidence.

The world of academia has been flooded with discussions as regards the approaches to carry out research. It should be noted that this discussion has brought to light an old dichotomy that has existed among researchers over a period of time, namely which of the approaches is superior to the other. Some scholars appear to support the quantitative research approach, while others would be more inclined towards the qualitative research approach. In both cases, analysis of research results indicate that the approach used to conduct the investigation should not be considered superior or inferior to the other per se, but it should be analyzed depending on their usefulness to answer the research questions pre-established for the investigation. It is possible that some researchers might not able to decide on the correct approach to be used due to the fact that his or her research questions have not been establishes specifically.

It is important to demystify the assumptions that the quantitative research approach is superior to the qualitative one. This means that both approaches need to be taken into account equally and be used according to needs of the investigation being done. To do so professors and scholars in charge of research courses in graduate colleges do not have to show their bias towards one or the other and allow students to choose freely between the quantitative and the qualitative traditions according to their needs and preferences. In some cases, it is true that

researchers and students choose one of the traditions mainly because of personal preferences, and this should be accepted as possible.

As it is explained by studies conducted in the area, personal preferences, likes and dislikes of research traditions should not be disregarded as an unimportant factor when analyzing research results and their success or failure. It is necessary to consider that when individuals prefer one research approach to the other and they are free to choose, they might obtain more successful results if they are comfortable with the task they are performing. In relation to this topic, it has been largely discussed that lower levels of anxiety when conducting research following the preferred tradition can bring useful insight into the affective area that influences research in the same way as the intellectual area does.

At this stage there is still not enough empirical evidence that researchers and students carrying out research in either tradition are more successful than the others but this does not mean that more research in the area should not be beneficial. It is interesting to notice that some surveys have demonstrated how the choice of research approach might be affected by gender, age and or ethnicity. It follows that among the considerations to be taken into account when studying this topic it should also be considered the fact that personal well-being and comfort play a considerable important role.

End of Essay

APPENDIX C:
DATA COLLECTION INSTRUCTION SHEET

EXPLANATION OF RESEARCH

Title of Project: Understanding ESL Teachers' Perception of Writing Quality

Principal Investigator: Alison Youngblood
Faculty Supervisor: Dr. Keith Folse; Dr. Joyce Nutta

The purpose of this research is to understand how ESL teachers rate the quality of **four (4) academic writing assignments**. All writing assignments have been written **by advanced non-native speakers**. All the essays were composed during a **30-minute in-class** writing activity. The students **did not have** any access to outside resources such as a dictionary, grammar text, or the Internet. These ESL students hope to get a **graduate degree** in the U.S.

Your participation in the study and any information you provide will be kept confidential. Your name or identifying information will not appear in the dissemination of results. You can withdraw at any time before, during, or after the data collection. There is no compensation provided for participation in this study. However, for every participant, the researcher will **donate $1** to Kids House which is a non-profit organization (501(c)3 Tax ID# 59-3415005) providing counseling and mental health services to child and adolescent second language learners in the Central Florida area.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints: Alison Youngblood, Doctoral Candidate, TESOL Ph.D. Program, College of Education and Human Performance, (407) 608-9296 or a.youngb@knights.ucf.edu; or Dr. Keith Folse, Faculty Supervisor, College of Arts and Humanities at (407) 823-4555 or Keith.Folse@ucf.edu; or Dr. Joyce Nutta, Faculty Supervisor, College of Education and Human Performance at (407) 823-4341 or Joyce.Nutta@ucf.edu.

IRB contact about your rights in the study or to report a complaint: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901.

INSTRUCTIONS

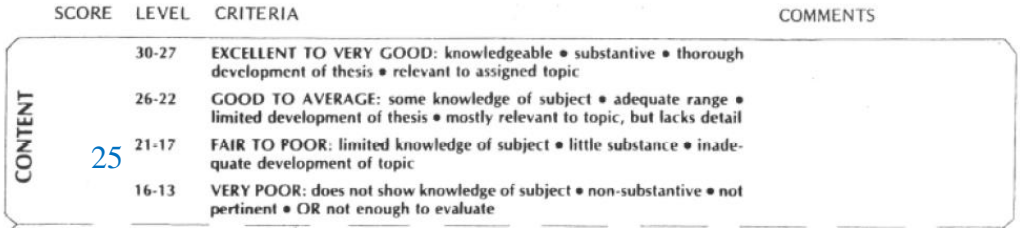
Included in this packet are four (4) academic essays and four (4) copies of a scoring rubric called the ESL Composition Profile. The ESL Composition Profile:

- Includes four sub-scales briefly summarized as:

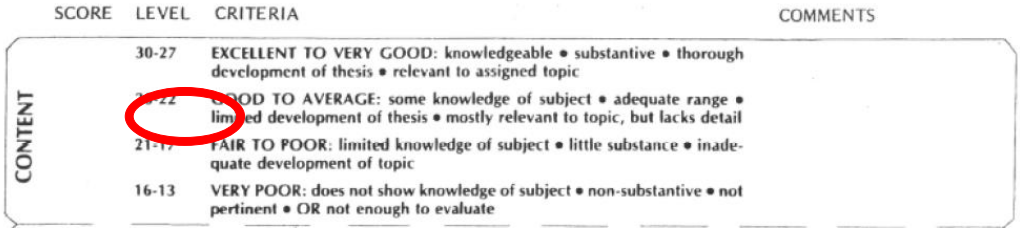
Content	*Clarity of the writer's message and idea development*
Organization	*Logical sequencing of ideas and use of cohesive devices*
Vocabulary	*Sophistication and effectiveness of the words*
Language Use	*Effective sentence structure, grammar, word order, and other conventions*
Mechanics	*Effective spelling, punctuation, and capitalization*

- Includes four score ranges for each sub-scale, and you can give a score of any **<u>whole number</u>** within each range: excellent to very good, good to average, fair to poor, very poor. **Write the score directly on the rubric**.

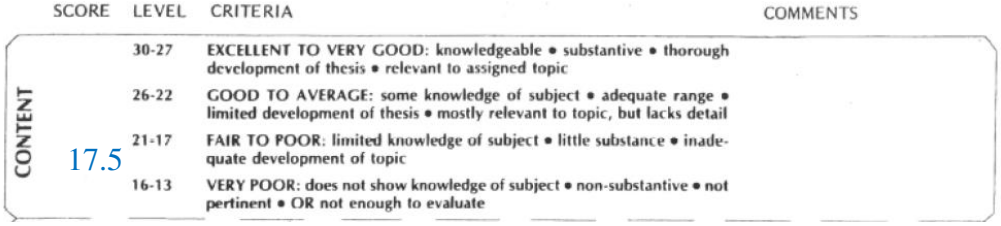
<u>CORRECT example:</u>



<u>INCORRECT example:</u>



<u>INCORRECT example:</u>



FOLLOW THESE STEPS TO COMPLETE THE RUBRIC:

1. Quickly **read** the essay. Based on your first impression of the essay, write a score for both the **content** and **organization** sub-scales.

2. Quickly **re-read** the composition and confirm your first impression of the essay. Score the remaining three sections: **vocabulary, language use,** and **mechanics.**

3. You do not have to include comments on the rubric.

THINGS TO REMEMBER:

- ✓ Grade quickly but with a purpose. Go with your gut!
- ✓ Work independently.
- ✓ The first essay may or may not be the lowest scoring essay. Keep an open mind!

THINGS TO AVOID

- ✓ Don't discuss your scores until all packets are collected
- ✓ Don't count errors in the paper.
- ✓ Don't estimate the score as you read.

APPENDIX D:
ESL COMPOSITION PROFILE

	SCORE	LEVEL	CRITERIA
CONTENT		30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
		26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail
		21-17	FAIR TO POOR: limited knowledge of subject • non-substantive • inadequate development of topic
		16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate
ORGANIZATION		20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive
		17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing
		13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development
		9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate
VOCABULARY		20-18	EXCELLENT TO VERY GOOD: sophistical range • effective word/idiom choice and usage • word form mastery • appropriate register
		17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*
		13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice usage • *meaning confused or obscured*
		9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate
LANGUAGE USE		25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
		21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*
		17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*
		10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate
MECHANICS		5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing
		4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*
		3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*
		2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate

APPENDIX E:
COPYRIGHT PERMISSION LETTER

CENGAGE Learning™

NELSON EDUCATION

Rights Administration and Content Reuse

20 Davis Drive, Belmont, California 94002 USA
Phone: 800-730-2214 or 650-413-7456 Fax: 800-730-2215 or 650-595-4603
Email: permissionrequest@cengage.com

Submit all requests online at www.cengage.com/permissions.

Request # **322614**

01/08/2014

Alison Youngblood
University of Central Florida
College of Education/School of Teaching and Learning
P.O. Box 161250
Orlando, FL 32816-1250 United States

Thank you for your interest in the following Cengage Learning/Nelson Education, or one of their respective subsidiaries, divisions or affiliates (collectively, "Cengage/Nelson") material.

Title:	TESTING ESL COMPOSITION 1E		
Author(s):	Jacobs 0883772256	ISBN:	9780838428993 (0838428991)
Publisher:	Heinle/ELT	Year:	1981
Specific material:	one-page rubric, or the ESL Composition Profile, as the grading instrument in my dissertation (page unknown)		
Total pages:	1		

For use by:

Name:	Youngblood
School/University/Company:	University of Central Florida
Course title/number:	dissertation study
Term of use:	Spring Term 2013

Intended use:
To copy or display for lecture or presentation, nonprofit research, training or counseling purposes use for which recipients are not charged. The number of copies may be changed to accommodate actual enrollment.

The non-exclusive permission granted in this letter extends only to material that is original to the aforementioned text. As the requestor, you will need to check all on-page credit references (as well as any other credit / acknowledgement section(s) in the front and/or back of the book) to identify all materials reprinted therein by permission of another source. Please give special consideration to all photos, figures, quotations, and any other material with a credit line attached. You are responsible for obtaining separate permission from the copyright holder for use of all such material. For your convenience, we may also identify here below some material for which you will need to obtain separate permission.

This credit line must appear on the first page of text selection and with each individual figure or photo:

> From Jacobs 0883772256. *TESTING ESL COMPOSITION*, 1E. © 1981 Heinle/ELT, a part of Cengage Learning, Inc. Reproduced by permission. www.cengage.com/permissions

Sincerely,

Donna Phillips
Permissions Associate

REFERENCES

Allen, R. (1966). Written English is a 'second language'. *Communication Studies, 18*(2), 81-85.

Altenburg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (Ed.), *Phraseology* (pp.101-122). Oxford, UK: Clarendon Press.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62,* 67-82.

Astika, G.G. (1993). Analytic assessments of foreign students' writing, *RELC Journal, 24*(1), 61-70. doi: 10.1177/003368829302400104

Bachman, L. (1990). *Fundamental considerations in language testing.* New York, NY: Oxford University Press.

Bauer, L., & Nation, I.S.P. (1993). Word families, *International Journal of Lexicography, 6*(4), 253-279.

Becker, J. (1975). The phrasal lexicon. In R. Shank, & B. L. Nash-Webber (Eds.). *Theoretical issues in natural language processing* (pp.60-63). Cambridge, MA: Bolt, Beranek, & Newman.

Bennet, G. R. (2010). *Using CORPORA in the language learning classroom: Corpus linguistics for teachers.* Ann Arbor, MI: University of Michigan Press.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*(2), 384-414.

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., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge, UK: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, UK: Longman.

Bloom, L.Z. (2006). Good enough writing: What is good enough writing, anyway? In P.Sullivan & H. Tinberg (Eds.), *What is "college-level" writing?* (pp.71-91). Urbana, IL: National Council of Teachers of English.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research, 10,* 245-261. doi: 10.1191/1362168806lr195oa

Bower, G. H. (1969). Chunks as interference units in free recall. *Journal of Verbal Learning and Verbal Behavior, 8*(5), 610-613. doi: 10.1016/S0022-5371

Brezina, V. & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the *New General Service List. Applied Linguistics Advanced Access.* Retrieved from http://applij.

Britton, J. (1975). *The development of writing abilities (11-18).* Urbana, IL: National Council of Teachers of English

Browne, C. (2013). The New General Service List: Celebration 60 years of vocabulary learning. *The Language Teacher, 37*(4), 13-16.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Canale, M., & Swain, M. (1981). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1-47.

Cayer, R. L. & Sacks, R. K. (1979). Oral and written discourse of basic writers: Similarities and differences. *Research in the Teaching of English, 13*(2), 121-128.

Chafe, W.& Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology, 16,* 383-407.

Cobb, T. (2013). Compleat lexical tutor. Retrieved from http://www.lextutor.ca/

Cohen, A. (2005). Teaching academic ESL writing: Practical techniques in vocabulary and grammar. *Studies in Second Language Acquisition, 27*, 109-110. doi: 10.1017/s0272263105240056

Conklin, K. & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32,* 45-61. doi: 10.1017/S0267190512000074

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397-423. doi: 10.1016/j.esp.2003.12.001

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17,* 391-406. doi: 10.1016/j.linged.

Council of Writing Program Administrators. (2008). *WPA outcomes statement for first-year composition.* Retrieved from wpacouncil.org/positions/outcomes.html

Cowie, A.P. (1998). Introduction. In A.P. Cowie (Ed.), *Phraseology* (pp. 1-20). Oxford, UK: Clarendon Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2012). Academic vocabulary, writing and English for academic purposes: Perspectives from second language learners. *RELC Journal, 43*(1)*,* 137-145. doi: 10.1177/0033688212439323

Coxhead, A., & Nation, I. S. P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew (Ed.). *Researching perspectives on English for academic purposes* (pp. 252-267). Cambridge, UK: Cambridge University Press.

Crossley, S. & Salsbury, T.L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics, 49,* 1-26. doi: 10.1515/iral.2011.001

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*(3), 573-605. doi: 10.1111/j.1467.9922.2010.00568.x

Crystal, D. (1995). *The Cambridge encyclopedia of the English language.* Cambridge, UK: Cambridge University Press.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question, and some other matters. *Working Papers on Bilingualism, 19*, 121-129.

Cummins, J. (1980). Psychological assessment of immigrant children: Logic or intuition? *Journal of Multilingual and Multicultural Development, 1,* 97-111.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics, 2,* 132-149.

Cummins, J. (2000). Putting language proficiency in its place: Responding to critiques of the conversational/academic language distinction. *English in Europe: The acquisition of a third language,* 54-83. Retrieved from http://iteachilearn.org/cummins/converacademlangdisti.html

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street, & N. H. Hornberger (Eds.). *Encyclopedia of language and education, 2nd edition, Volume 2: Literacy.* New York, NY: Springer Science and Business Media.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computer, 25*(4), 447-464. doi: 10.1093/llc/fqq018

Davis, M. (2008). *The corpus of contemporary American English: 450 million words, 1990-present.* Retrieved online at http://corpus.byu.edu/coca/

Dechert, H. (1984). Individual variation in language. In H. Dechert, D. Möhle, & M. Raupach (Eds.), *Second language productions* (pp.156-185). Tübingen, Germany: Gunter Narr.

DeVito, J. (1967). A linguistic analysis of spoken and written language. *Communication Studies, 18*(2), 81-85.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method.* Hoboken, NJ: Wiley & Sons, Inc.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics, 47,* 157-177. doi: 10.1515/iral.2009.007.

Ellis, N. (1998). Emergentism, connectionism, and language learning. *Language Learning, 48*, 631-664.

Ellis, N. (1999). Cognitive approaches to SLA. *Annual Review of Applied Linguistics, 19* 22-42.

Ellis, N. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. Long (eds.). *Handbook of second language acquisition*, (pp. 63-103). Malden, MA: Blackwell.

Ellis, N. (2005). At the interface: dynamic interactions of explicit and implicit knowledge. *Studies in Second Language Acquisition, 27*(2), 305-352.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61-78. doi: 10.1515/CLLT.2009.003

Ellis, N.C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375-396.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford, UK: Oxford University Press.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication, 28*(2), 122-128.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Fang, Z., & Schleppegrell, M. J. (2010). Disciplinary literacies across content areas: Supporting secondary reading through functional language analysis. *Journal of Adolescent & Adult Literacy, 53*(7), 587-597. doi: 10.1598/JAAL.53.7.6

Ferris, D. R. (1994). Lexical and syntactical features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Ferris, D. R. (2009). *Teaching college writing to diverse student populations.* Ann Arbor, MI: University of Michigan Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In F. R. Palmer (Ed.). *Studies in linguistic analysis: Special volume of the Philological Society* (pp. 168-205). Oxford, UK: Oxford University Press.

Folse, K. (2010). Is explicit vocabulary focus the reading teacher's job? *Reading in a Foreign Language, 22*(1), 139-160.

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching.* Ann Arbor, MI: University of Michigan Press.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). Boston, MA: McGraw-Hill Higher Education

Furneaux, C., Paran, A., & Fairfax, B. (2007) Teacher stance as reflected in feedback on student writing: An empirical study of second language school teachers in five countries. *International Review of Applied Linguistics, 45*(1), 69-94.

Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics.* Advanced online publication. doi: 10.1093/applin/amt015

Gee, J. P. (1990). *Social linguistics and literacies: Ideologies in discourses.* New York, NY: Falmer Press.

Gibbons, P. (1991). *Learning to learn in a second language.* Newtown, AU: Primary English Teaching Association.

Glass, G. V. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288.

Goldenberg, C. (2008). Teaching English language learners: What the research does- and does not- say. *American Educator, 32*(2), 11-23, 42-43.

Gonzalez, M, Youngblood, A., & Giltner, E. (2012). Student-initiated linguistic-based feedback versus process-oriented feedback in foreign language writing. *Florida Foreign Language Journal, 9*(1), 11-22.

Gonzelez, M. C. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker academic compositions* (Unpublished doctoral dissertation). University of Central Florida, Orlando, Florida.

Graesser, A. C., McNamara, D. S. Louwerse, M. M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*(2), 193-202.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (eds.). *Phraseology: An interdisciplinary perspective* (pp. 27-49). Amsterdam: John Benjamins

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (Ed.), *Phraseology* (pp. 145-160). Oxford, UK: Clarendon Press

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*(1), 75-85.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218-238. doi: 10.1016/j.asw.2013.05.002

Halliday, M. A. K. (2004). Lexicology. In M. A. K. Halliday, W. Teubert, C. Yallop, & A. Čermáková (Eds.). *Lexicology and corpus linguistics* (pp. 1-22). London, UK: Continuum.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Harlow, UK: Longman.

Halliday, M.A.K. (1979). Differences between spoken and written language: Some implications for literacy teaching. *Proceedings of the 4th Australian Conference, Australia, 2,* 37-52.

Harwood, N. (2002). Taking a lexical approach to teaching: Principles and problems. *Journal of Applied Linguistics, 12*(2), 139-155.

Hewings, M. & Hewings, A. (2002). "It is interesting to note that…": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes, 21,* 367-383.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London, UK: Routledge.

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.). *Phraseology* (pp. 161-186). Oxford, UK: Claredon.

Hu, M. & Nation, I.S.P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Hunt, K. W. (1964). *Grammatical structures written at three grade levels* (Report No. 3). Urbana, IL: National Council of Teachers of English.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow, UK: Longman.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Hyland, K., & Tse, P. (2009) Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies, 9*(2), 111-129.

International WAC/WID Mapping Project. (2008). *2006-2008 national survey of US WAC/WID initiatives.* Retrieved from mappingproject.ucdavis.edu/

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, MA: Newbury House.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (ed.). *Formulaic sequences* (pp. 269-292). Amsterdam: John Benjamins.

Juliand, A. & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words.* The Hague: Mouton.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*(4), 322-332. doi: 10.1093/elt/ccio61

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal, 73*(4), 440-464.

Krejce, R. (1970). Determining sample size for research activities. *Educational and Psychological Measurement, 30,* 607-610.

Kuiper, K. (1996). *Smooth talkers: The linguistic performance of auctioneers and sportscasters.* Mahwah, NJ: Lawrence Erlbaum.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27*(1), 123-134.

Laufer, B. & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In Coady & T. Huckin (Eds.), *Second Language Vocabulary Acquisition* (pp. 20-34). Cambridge, England: Cambridge University Press.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size tests of controlled productive ability. *Language Testing, 16*(1), 33-51.

Lee, C. D., & Spratley, A. (2010). *Reading in the disciplines: The challenges of adolescent literacy.* New York, NY: Carnegie Corporation of New York.

Leki, I., &Carson, J. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly, 28*(1), 81-101.

Lervåg, A. & Aukrust, V. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *The Journal of Child Psychology and Psychiatry, 51*(5), 612-620. doi: 10.111/j.1469-7610.2009.02185.x

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18,* 85-102. doi: 10.1016/j.jslw.2009.02.001

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. doi: 10.1093/applin/ams010

McCarthy, M. (1998). *Spoken language and applied linguistics.* Cambridge: Cambridge University Press.

Mcleod, S. H. (1992). Writing across the curriculum: An introduction. In S. H. McLeod, & M. Soven (Eds.). *Writing across the curriculum* (pp. 1-8). Thousand Oaks, CA: Sage

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language teaching and Linguistics Abstracts, 13*(3-4)*,* 221-246. doi: 10.1017/S0261444800008879

Michigan corpus of upper-level student papers. (2009). Ann Arbor, MI: The Regents of the University of Michigan.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Moon, R. (1998). Frequencies and forms of phrasal lexemes in English. In A.P. Cowie (Ed.), *Phraseology* (pp. 79-100). Oxford, UK: Clarendon Press.

Nagy, W. E., & Anderson, R. C. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*(2), 237-270. doi: 10.3012/00028312024002237

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly,* 24(3), 262-282.

Nation, I. S. P. (2001a). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. M. Koski, R. D. Sell, & B. Warvik (Eds.). *Language, learning, and literature: Studies presented to Hakan Ringbom* (pp. 167-181). Turku, Finland: Abu Akademi University English Department.

Nation, I. S. P. (2001b). *Learning Vocabulary in Another Language*. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Nation, I. S. P., & Webb, S. (2010). *Researching and analyzing vocabulary.* Boston, MA: Heinle Cengage.

National Clearinghouse for English Language Acquisition. (2011). *Nation's report card: Mathematics 2011* (NCES 2012-458). Washington, D. C.: U.S. Department of Education, Institute of Educational Sciences.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching.* Oxford, UK: Oxford University Press.

Nesselhauf, N., &Tschichold, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning, 15*(3), 251-279. doi: 10.1076/call.15.3.251.8190

Nutta, J., Mokhtari, K., & Strebel, C. (2012). *Preparing every teacher to reach English learners.* Cambridge, MA: Harvard Education Press.

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18*(1), 83-108. doi: 10.1075/ijcl.18.1.070do

Oakes, M. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Ogden's Basic English (2012, March 24). Retrieved from http://ogden.basic-english.org/basiceng.html

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcskik, H. Ouali, & K. M. Wheatley (Eds.). *Formulaic Language Volume 2: Acquisition, loss, psychological reality, and functional explanations* (pp. 375-386). Amsterdam: John Benjamins.

Oppenheim, N. (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In H. Riggenbach (Ed.). *Perspectives on fluency* (pp. 220-240). Ann Arbor, MI: University of Michigan Press.

Pawley, A. (1991). How to talk cricket: On linguistic competence in a subject matter. In R. Blust (Ed.). *Currents in Pacific linguistics: Papers on Austronesian languages and*

ethnolignuistics in honour of George W. Grace (pp. 339-368). Canberra, AU: Pacific Linguistics.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.). *Language and communication* (pp. 191-225). London, UK: Longman.

Perkins, K. (1980). Using objective measures of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14*(1)*,* 61-69.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 70-91.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 1-63). Cambridge, UK: Cambridge University Press.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48,* 281-317.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329-363. doi: 10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* London, UK: Palgrave Macmillan.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.). *Formulaic sequences: Acquisition, processing, and use* (pp. 1-22). Amsterdam: John Benjamins Publishing

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145-171.

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.). *Formulaic sequences: Acquisition, processing, and use* (pp. 55-86). Amsterdam: John Benjamins Publishing Company.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40-59.

Simon, H. A. (1974). How big is a chunk? *Science, 183,* 482-488.

Simpson-Vlach, R., & Ellis, N. C. (2010) An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. doi: 10.1093/applin/amp058

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Spear, M. (1997). Controversy and consensus in freshman writing: An overview of the field, *Review of Higher Education, 20*(3), 319-322.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: does the type of target language influence the association? *IRAL, 49,* 321-343. doi: 10.1515/iral.2011.017

Strunk, W., & White, E.B. (2000). *Elements of Style* (4th ed.). Boston, MA: Allyn and Bacon.

Stubbs, M. (1995). Corpus evidence for norms of lexical collocation. In G. Cook & B. Seidlhofer (Eds.). *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 245-256). Oxford, UK: Oxford University Press.

Sullivan, P. (2006). An essential question: What is "college-level" writing?. In P. Sullivan & H. Tinberg (Eds.). *What is "college-level" writing?* (pp. 1-30). Urbana, IL: National Council of Teachers of English.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

Thorndike, E. L (1921). *The teacher's word book*. New York: Teachers College, Columbia University Press.

Tsui, A. (2004). What teachers have always wanted to know-and how corpora can help. In J. Sinclair (Ed.). *How to use corpora in language teaching* (pp.39-61). Amsterdam: John Benjamins.

Vrooman, A. H. (1967). *Good writing: An informal manual of style.* New York, NY: Antheneum.

West, M. (1953). *A general service list of English words.* London, UK: Longman.

White, E. M. (2007). *Assigning, responding, evaluating: A writing teacher's guide* (4th ed.). Boston, MA: Bedford/St. Martin's

Wilkins, D. (1972). *Linguistics in language teaching*. London, UK: Edward Arnold.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *The Canadian Modern Language Review, 63*(1), 13-33.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam & L. K. Obler (Eds.). *Bilingualism across the lifespan: Aspects of acquisition, maturity and loss* (pp. 55-72). Cambridge, UK: Cambridge University Press.