

University of Central Florida

STARS

Electronic Theses and Dissertations

2018

Analyses and Comparisons of Three Lexical Features in Native and Nonnative Academic English Writing

Xiaoli Yu

University of Central Florida



Part of the [Language and Literacy Education Commons](#)

Find similar works at: <https://stars.library.ucf.edu/etd>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Yu, Xiaoli, "Analyses and Comparisons of Three Lexical Features in Native and Nonnative Academic English Writing" (2018). *Electronic Theses and Dissertations*. 6061.

<https://stars.library.ucf.edu/etd/6061>

ANALYSES AND COMPARISONS OF THREE LEXICAL
FEATURES IN NATIVE AND NONNATIVE ACADEMIC
ENGLISH WRITING

by

XIAOLI YU

B.A. Shandong University of Economics and Finance, 2012
M.A. University of Kansas, 2014

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

Summer Term

2018

Major Professor: Keith S. Folse

© 2018 Xiaoli Yu

ABSTRACT

Built upon the Contrastive Interlanguage Analysis (CIA) framework, this corpus-based research analyzes three lexical features (lexical diversity, lexical sophistication, and cohesion) in native and nonnative English writers' academic writing and examines the potential differences in lexical performance 1) between native and nonnative English writers and 2) across all writers from various language backgrounds. The differences in lexical performance in academic writing between native and nonnative English writers and the unique characteristics of writers from different language backgrounds suggest the necessity of targeted academic writing instruction based upon learner needs. Using text length as the covariate, two Multivariate Analysis of Covariate (MANCOVA) were conducted with language background as the Independent Variable and the three lexical features as the Dependent Variables. The results revealed that nonnative English writers demonstrated significantly lower performance in lexical sophistication than did native English writers. In terms of the comparison between writers from different language backgrounds, the results suggested statistically significant differences in all three aspects of lexical features. Pedagogical implications for vocabulary instruction in academic writing for nonnative English writers include emphasizing the mastery of academic, low-frequency, and discipline-specific vocabulary. In addition, improving nonnative writers' vocabulary size and lexical diversity can offer these learners more options to build cohesion in academic writing at a deeper level. Moreover, the results of this study highlight the wide but often under-considered variability within any language group as individual learner differences come into play, thereby downplaying the idea that writers of any given language tend to perform homogeneously. Instructors should acknowledge the unique writing characteristics of different nonnative writers

and their varied learner needs. Thus, targeted instruction is essential to provide effective enhancement to nonnative English writers' lexical performance in academic writing.

Keywords: lexical features, academic writing, learner corpus research

To the triumph of love over long distance

ACKNOWLEDGMENTS

This dissertation, as well as the successful completion of the past three years of my life, would not have been possible without the support of my mentors, family, friends, and colleagues. The following simple words of gratitude are offered from the bottom of my heart to express my appreciation to those people.

First and foremost, my deepest gratitude goes to Dr. Keith Folse for his continuous guidance, support, and encouragement ever since I started this Ph.D. program. Throughout the journey of completing this dissertation, Dr. Folse has spent countless hours on locating suitable resources, meeting with me, revising drafts, and answering emails. Without his absolute faith in my capabilities, I could not possibly have reached this stage of my doctoral work. Also, Dr. Folse's dedication to the field has taught me a great deal about being an educator, scholar, and mentor.

I would also like to thank Dr. David Boote, who provided inspiring advice on the research design and statistical analysis of my study. I am very much grateful to have Dr. Boote on my dissertation committee because he spent seemingly unlimited personal time on reading research articles about applied linguistics and corpus studies, as well as giving sincere suggestions to my writing. I am fortunate to have had this opportunity to learn from his direction.

I also owe a huge debt of gratitude to the other two committee members, Dr. Beth Young and Dr. Francisco Fernandez-Rubiera. Dr. Young's knowledge of corpus studies and her constructive guidance aided me in solving the research design problems from more practical perspectives. Likewise, I am very much grateful to Dr. Fernandez-Rubiera for his insights into applied linguistics as well as his encouragement throughout this dissertation process.

Next, I would like to express my sincere gratitude to all my professors and colleagues in the TESOL Ph.D. program at the University of Central Florida. I truly appreciate their efforts in organizing the program and providing mentorship to students. My sincere appreciation also goes to all my colleagues and friends in the program for helping me balance my personal life and academic study. Without the solid establishment of this program, I could not have dreamed of completing my degree so smoothly.

My most earnest thanks are for my best friend and lifetime partner, Veysel Altunel. Without his encouragement, I would not have started this journey in the first place. It is his faithful companionship throughout the past three years that helped us conquer the distance of six thousand miles.

Final and special thanks are reserved for my parents, grandparents, and other family members who have raised me well and taught me how to be a good person. To my grandpa, thank you for the unswerving confidence in me for my entire life; to my grandma, thank you for the unconditional love to me in anytime; to my mom, thank you for being the best mom I can ever dream of; and to my dad, thank you for giving me the opportunity to make you proud. No venture of this magnitude is ever a single person's trip, and I thank everyone for joining me on this journey.

最后特别感恩父母、祖父母和所有从小到大爱护我的家人。你们无微不至的爱让我能得以健康长大成人，你们身体力行让我知道做事之前先如何做人。感恩姥爷对我永远的支持与鼓励；感恩姥姥从小到大给予我所有的无条件的关心和爱；感恩妈妈一直是我心中最美好的存在；感恩爸爸让我有机会让你自豪。这三年的读博历程不是我一个人在走，感恩陪伴我完成这段人生经历的每一个人。

TABLE OF CONTENTS

LIST OF FIGURES	xiv
LIST OF TABLES	xv
CHAPTER ONE: INTRODUCTION.....	1
Background	4
Rationale.....	6
Research Questions	8
Hypotheses	8
Hypotheses for Research Question One.	8
Hypotheses for Research Question Two	9
Significance of the Study	10
Study Limitations	12
Definitions of Terms for the Study	13
Organization of the Study	14
CHAPTER TWO: RESEARCH AND LITERATURE REVIEW	16
The Nature of Academic Language	18
The Role of Vocabulary in SLA and TESOL	19
Vocabulary and Listening, Reading, and Speaking.....	20
Vocabulary and Writing	23
Summary.....	30

Theoretical and Methodological Foundation of the Present Study	30
Corpus Linguistics, SLA, and TESOL	30
Learner Corpus Research (LCR) and Contrastive Interlanguage Analysis (CIA)	37
Summary.....	48
Lexical Features and Corresponding Measurements	49
Major Concepts and Definitions in Corpus Linguistics	50
Major Lexical Features in Writing	52
Target Lexical Features and Measurements	57
Summary.....	72
Conclusion.....	72
CHAPTER THREE: RESEARCH METHODOLOGY	74
Research Questions	75
Hypotheses	75
Hypotheses for Research Question One	75
Hypotheses for Research Question Two	76
Orientation to Research Design.....	77
Target Population and Selected Corpora.....	79
The International Corpus of Learner English (ICLE): Nonnative English Speakers	80
The Louvain Corpus of Native Essay Writing (LOCNESS): Native English Speakers	88

Sample Size and Sampling Procedures	93
Sample Size Determinations.....	93
Sampling Procedures	103
Data Extraction and Measurements.....	103
Lexical Diversity	104
Lexical Sophistication	104
Cohesion	105
Conclusion.....	107
CHAPTER FOUR: RESULTS	108
Research Questions	108
Sampling Procedures.....	109
Descriptive Data Results	111
Initial Data Screening of the Independent Variable	115
Initial Data Screening of the Dependent Variables	115
Assumption Tests and Final Data Screening.....	124
Research Questions One.....	127
Research Questions Two.....	130
Post Hoc Analyses.....	132
Lexical Diversity	133

Coverage of the First 1,000 Words.....	134
Coverage of the Second 1,000 Words	135
Coverage of the High-Frequency Words.....	136
Coverage of the AWL.....	137
Coverage of the AVL	138
Coverage of Off-list Words	140
Referential Cohesion	141
LSA.....	142
Connectives	144
Conclusion.....	148
CHAPTER FIVE: DISCUSSION.....	150
Purpose of the Study	150
Summary of the Findings	153
Research Questions One	155
Research Questions Two	157
Significance of the Findings.....	158
Limitation of the Study	161
Pedagogical Implications	162
Recommendations for Further Research	169

Conclusion.....	171
APPENDIX A: IRB LETTER OF APPROVAL	172
APPENDIX B: COPYRIGHT PERMISSION LETTER	174
APPENDIX C: COPYRIGHT PERMISSION LETTER	176
APPENDIX D: COPYRIGHT PERMISSION LETTER	178
LIST OF REFERENCES	180

LIST OF FIGURES

Figure 1: Contrastive Interlanguage Analysis (CIA)	43
Figure 2: CIA ²	47
Figure 3: Lexical Richness.....	55
Figure 4: Research Design	79
Figure 5: Task and Learner Variables of the ICLE.....	82
Figure 6: Mean Differences in Lexical Diversity across Writers from Various Language Groups	146
Figure 7: Mean Differences in Lexical Sophistication across Writers from Various Language Backgrounds	147
Figure 8: Mean Differences in Cohesion across Writers from Various Language Backgrounds	147

LIST OF TABLES

Table 1 Vocabulary and Listening.....	21
Table 2 Vocabulary and Reading.....	22
Table 3 Vocabulary and Speaking	23
Table 4 Vocabulary and Writing.....	28
Table 5 Surface Level Measures of Lexical Features (Polio, 2001).....	54
Table 6 Lexical Diversity and Holistic Writing Quality.....	59
Table 7 Measurements of Lexical Diversity	61
Table 8 Lexical Sophistication and Holistic Writing Quality.....	63
Table 9 Cohesion and Holistic Writing Quality	70
Table 10 The ICLE and the LOCNESS	78
Table 11 Top 10 Topics in the ICLE	84
Table 12 Task Variables in the ICLE	84
Table 13 CEFR Results - 20 Essays Per Subcorpus	86
Table 14 Subcorpora Size in the ICLE	87
Table 15 Subcorpora Size in the LOCNESS	89
Table 16 Essays in the LOCNESS.....	90
Table 17 Publications Based on the ICLE and the LOCNESS.....	92
Table 18 Summary of the Target Measurements of Lexical Features	93
Table 19 Measurements of Cohesion.....	98
Table 20 Three Factors Extracted from the EFA.....	101
Table 21 Two Types of MANCOVA	102

Table 22 Summary of the Dependent Variables	106
Table 23 Number of Essays by Language Designation	110
Table 24 Text Length.....	112
Table 25 Descriptive Statistics for Lexical Features	114
Table 26 Multivariate Outliers Identified Based on Mahalanobis Distance.....	117
Table 27 Number of Essays after Removing Outliers Based on Mahalanobis Distance.....	118
Table 28 Skewness of Dependent Variables.....	119
Table 29 Kurtosis of Dependent Variables.....	119
Table 30 Outliers Identified Based on Skewness and Kurtosis	121
Table 31 Number of Essays after Removing Outliers based on Skewness and Kurtosis	122
Table 32 Final Descriptive Statistics of Skewness and Kurtosis.....	123
Table 33 Levene's Test with Two Language Groups (Native and Nonnative)	125
Table 34 Levene's Test with Seven Language Groups	126
Table 35 Correlation (Pearson's r) between Dependent Variables.....	127
Table 36 Tests of Between-Subjects Effects for Measures of Lexical Sophistication (MANCOVA 1)	129
Table 37 Tests of Between-Subjects Effects for Lexical Diversity and Measures of Cohesion (MANCOVA 1)	129
Table 38 Tests of Between-Subjects Effects for All Lexical Features (MANCOVA 2).....	132
Table 39 Descriptive Statistics of Lexical Diversity	134
Table 40 Pairwise Comparisons (Mean Difference) of Lexical Diversity	134
Table 41 Descriptive Statistics of the Coverage of the First 1,000 Words.....	135
Table 42 Pairwise Comparisons (Mean Difference) of the Coverage of the First 1,000 Words	135

Table 43 Descriptive Statistics of the Coverage of the Second 1,000 Words	136
Table 44 Pairwise Comparisons (Mean Difference) of the Coverage of the Second 1,000 Words	136
Table 45 Descriptive Statistics of the Coverage of the First 2,000 Words.....	137
Table 46 Descriptive Statistics of the Coverage of the AWL.....	138
Table 47 Pairwise Comparisons (Mean Difference) of the Coverage of the AWL.....	138
Table 48 Descriptive Statistics of the Coverage of the AVL.....	139
Table 49 Pairwise Comparisons (Mean Difference) of the Coverage of the AVL.....	140
Table 50 Descriptive Statistics of the Coverage of Off-list Words	140
Table 51 Pairwise Comparisons (Mean Difference) of the Coverage of Off-list Words	141
Table 52 Descriptive Statistics of Referential Cohesion	142
Table 53 Pairwise Comparisons (Mean Difference) of Referential Cohesion	142
Table 54 Descriptive Statistics of the LSA.....	143
Table 55 Pairwise Comparisons (Mean Difference) of the LSA.....	144
Table 56 Descriptive Statistics of Connectives	145
Table 57 Pairwise Comparisons (Mean Difference) of Connectives.....	145

CHAPTER ONE: INTRODUCTION

University student populations are changing. Since the turn of the last century, the demographics in educational institutions around the globe are different than in years past. In particular, the population of students traveling to other countries for tertiary-level studies doubled in 2012 compared to the year 2000, reaching four million in total (UNESCO Institute for Statistics, 2016). In the United States, 5% of the 20 million students enrolled in higher education institutions in the 2015-2016 academic year were international students from non-English language backgrounds (Witherell & Department of State, 2016).

Logically, this more diverse population of students requires an adjustment in how learning can take place, an adjustment by both those who teach and those who learn. Many institutions, for example, have implemented programs with appropriate strategies among their faculty and staff to increase awareness of this diversity. At the same time, students who are nonnative speakers (NNSs) studying in native-speaking countries must recognize potential challenges and make efforts to meet the requirements from their host institutions in order to succeed in academics.

Among the various difficulties that NNSs need to overcome, language, especially academic language, is the fundamental hurdle. NNSs tend to find their lack of proficiency in the target language impedes their progress in academic studies. As a direct result, educators with classes of NNSs also recognize the challenge of providing effective instruction to enhance these students' academic language proficiency.

Language learning is not a single-dimensional topic. In fact, four skills, including listening, speaking, reading, and writing, are all essential components of language proficiency. To do well, university students should have solid ability in all four key language skills. However,

in the environment of higher education, perhaps the most challenging but inevitable skill that all students need to develop is writing.

In English-speaking environments, for both native and nonnative English speakers (NESs & NNEs), presenting their learning in written format is almost a compulsory activity in most English-medium institutions in higher education. Research papers, journals, and reflections are a few examples for writing assignments. In university lectures, students may not have the opportunity to express their opinions by speaking in class; and in many courses, a student can rarely speak up, if ever, and still earn a grade of A. In contrast, almost all students are required to complete writing assignments. In particular, for NNEs, before attending English-medium institutions, they are required to take standardized examinations to demonstrate their English language proficiency. Writing is always an essential section in these high-stakes examinations. Therefore, based upon the significance of writing in achieving success in higher education, the current study focuses on issues related to improving writing performance for NNEs.

To address the issue of writing and identify effective pedagogical strategies, it is essential to ask what the integral factors of good academic writing are. In answering this question, a large body of scholarship has suggested the centrality of vocabulary in achieving better performance in academic writing. Gardner (2013), for instance, considers vocabulary as the fuel that motivates the moving of communication, especially in written format. Moreover, from learners' perspective, utilizing appropriate vocabulary in academic writing is a major concern (Coxhead, 2012). However, vocabulary instruction has yet earned a sufficient amount of time and efforts in writing classes nor other university courses (Meara, 2002). Thus, the present study focuses on vocabulary and lexical features in academic writing of NNEs, the NESs are used as the referential population.

The development of pedagogical strategies should be built upon actual learner needs. To better understand learner needs, analyzing the language that learners produce is a promising path because it allows us to note the differences between the writing of diverse NNSs and the ideal academic writing. Through analyzing the linguistic features of learner language, researchers are able to identify the strengths and weaknesses of the language and thus propose appropriate instructional methods. Consequently, researching lexical features of interlanguage products written by English learners (ELs) is a critical aspect of the present study.

Instructors often expect the NNSs to have a similar writing performance to that of NSs. Taking into account of the interlanguage from ELs' writing, the current study analyzes lexical characteristics of NNEs' writing samples and compares them to the NESs' products to reveal the potential differences between the two groups of writers. This analysis is beneficial for helping NNEs and their instructors recognize the weaknesses that NNEs might have in improving lexical quality in their academic English writing.

While comparing the writing of NSs and NNSs may seem somewhat straightforward, the population of ELs is difficult to be generalized. Any group of ELs already represents a potentially huge contrast. Various mother tongue backgrounds, educational experiences, and proficiency levels encompass different features of their interlanguages. As a result, comparing linguistic features within the NNEs group is also a key question. The influence of first language (L1) on second language (L2) acquisition has been discussed extensively. Especially in writing, L1 transfer is considered as a critical factor that influences one's writing performance (Berman, 1994). Therefore, another important component of this study is to compare lexical features in academic writing samples across all writers from different mother tongue backgrounds and capture any potential influences from L1.

In sum, the research purpose of the present study is to examine the differences in lexical features in academic English writing 1) between NNEs and NESs and 2) across all speakers from various language backgrounds. By demonstrating the differences, the present study aims to help educators, textbook writers, and curriculum designers to understand 1) the differences between NNEs and their native speaking peers and 2) the variability among NNEs.

This chapter first elucidates the general background knowledge of lexical component in academic writing. Next, the reasons why the research objectives in the present study are important for investigation are presented as the rationale section, which is followed by the research questions and corresponding hypotheses. Then, the limitations of the research are addressed. Key terminologies and their definitions are listed in the next section. Lastly, the chapter concludes with the organization of the remaining dissertation.

Background

It is a complex process from initially encountering a word to successfully using it in writing. Correct understanding, spelling, collocations, and grammatical features of the word are all essential components of using it appropriately (Coxhead & Byrd, 2007). For second or foreign language learners, limited vocabulary knowledge often leads to the gap between what they want to convey and what they can convey (Laufer, 2013).

Empirical studies have suggested the close correlation between vocabulary knowledge and L2 writing performance. For instance, Laufer and Nation (1995) emphasize the important influence of vocabulary size and lexical richness on L2 writing quality. Hyland (2007) notes the critical role of vocabulary depth knowledge in writing appropriately based upon genre and disciplinary features. Studies have also suggested the relationship between vocabulary and writing from learners' perspective. Zhou (2009) and Coxhead (2012) both explored learners'

perceptions of vocabulary in academic writing, noting that learners are conscious of the crucial role of vocabulary in improving writing performance. Meanwhile, the studies also demonstrate apparent demands of sufficient and effective vocabulary instruction from the learners.

Some commonly researched areas of vocabulary in writing include vocabulary size, vocabulary depth of knowledge, and various lexical features in writing. Vocabulary size and depth of knowledge are generally analyzed via receptive and productive tests. However, the results of these tests may differ depending on the nature of the tests and participants. In addition, the number of participants recruited in the studies can be extremely limited. In contrast, different from conducting the vocabulary tests, analyzing lexical features in written discourses can be based on adequate texts and more advanced examining techniques.

The analysis of language usage based upon a large amount of authentic and naturally-occurred texts is referred to as corpus-based analysis (McEnery & Wilson, 2001). The use of corpus-based data provides an empirical basis for determining lexical features of a specific group of writers. The commonly researched aspects of lexis in writing include lexical richness, lexical diversity, lexical errors, and so forth. The definitions and classification of these key terms are introduced in detail in Chapter Two.

The present study explores learners' needs by analyzing lexical features in their writing. The corpora selected in this study contain sufficient and authentic learner English in writing, which establishes a solid foundation for analyzing various lexical features in learner written language. Therefore, the strengths and weaknesses of learner writing can be analyzed through empirical evidence rather than what we might imagine "good writing" to be.

Rationale

As mentioned in prior sections, writing is inevitable in higher education. The writing-across-the-curriculum (WAC) movement further promotes involving writing as a major component of assessment in college-level classes (Britton, 1975). At the University of Central Florida, the WAC program collaborates with faculty from all disciplines to develop theoretically and pedagogically sound and sustainable models of writing instruction across the curriculum (Writing Across the Curriculum, 2018). Workshops and consultations are regularly provided to assist faculty members from different disciplines. This shows the emphasis placed on writing from the university's perspective. However, investigations of NNEs suggest that most learners find writing as the most challenging task in completing a college class (Burke & Wyatt-Smith, 1996). Moreover, research in second language acquisition (SLA) has addressed the slow and complicated process for NNSs to achieve the desired performance in academic writing. Thus, investigating academic writing and developing more effective instructional strategies for NNSs are of great value to universities with growing numbers of international and nonnative speaking students.

Among the many diverse aspects that contribute to ultimate achievement in writing, including vocabulary, grammar, and structure, the use of vocabulary establishes the most fundamental quality of a composition (Laufer, 2013). Lexical variety, richness, sophistication, and errors are a few representative features of lexis that closely correlate with the holistic writing performance. The present study focuses on analyzing lexical aspects in academic writing to provide empirical reasons of why NNEs' writing is lexically different from that of their native speaking peers.

The current study is a non-experimental empirical research study, which employs a comparative method to look at the two-level differences in lexical features in academic writing: 1) between NNEs and NESs; 2) across all writers from various language backgrounds. First, the analyses of lexical features provide a comprehensive picture of learner English (i.e., interlanguage) in terms of academic writing. Second, having a referential variety (i.e., NESs) is beneficial for developing instructional strategies with an initial objective. Hence, the comparative research design is appropriate for conducting the current study.

In order to provide an accurate and representative picture of lexis in learner language and the differences between NESs and NNEs, corpus-based data is employed in the present study. The compositions in the corpora were collected from naturally-occurred student writing samples. With a large amount of written texts, the corpora provide an authentic and solid basis for empirical analysis of any language; enriched description of the language can be obtained through advanced computational techniques (Sinclair, 1996; St John, 2001; Zanettin, 1994).

Lastly, studies have investigated lexical features of NNEs' writing from different perspectives. For instance, Laufer and Nation (1995) focused on the feature of lexical richness; Chandler (2003) investigated lexical errors and the corresponding feedback; Jarvis (2002) addressed the issue of lexical diversity. However, very few of the empirical studies have analyzed the issue in a comprehensive matter. To better understand the lexical features of learner writing, it is essential to examine the issue from a thorough perspective. Thus, the present study investigates lexical features of native and nonnative speakers' academic writing and compare their differences in three major aspects, including lexical diversity, lexical sophistication, and cohesion. Detailed explanations of these features are presented in Chapter Two.

In sum, the importance of writing in academics and the significance of vocabulary in academic writing determine the specific research area of the present study. Following the research objectives, analyzing NNEs' lexical features in academic writing provides informative evidence for educators to develop targeted instructional methods based on learner characteristics and conduct needs analysis according to the empirical findings. Thus, the present study employs corpus-based data to present 1) the potential differences in lexical quality between NNEs' and NESs' academic writing and 2) the possible diversity across all writers from various language backgrounds.

Research Questions

The current study aims to answer the following research questions:

1. Are there significant differences in lexical features between native and nonnative academic English writing, as measured by lexical diversity, lexical sophistication, and cohesion?
2. Are there significant differences in lexical features, as measured by lexical diversity, lexical sophistication, and cohesion, in academic English writing across all writers from various mother tongue backgrounds?

Hypotheses

Hypotheses for Research Question One

Lexical diversity

H₀: There are no significant differences in the level of lexical diversity (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of lexical diversity (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Lexical sophistication

H₀: There are no significant differences in the level of lexical sophistication (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of lexical sophistication (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Cohesion

H₀: There are no significant differences in the level of cohesion (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of cohesion (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Hypotheses for Research Question Two

Lexical diversity

H₀: There are no significant differences in the levels of lexical diversity (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of lexical diversity (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Lexical sophistication

H₀: There are no significant differences in the levels of lexical sophistication (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of lexical sophistication (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Cohesion

H₀: There are no significant differences in the levels of cohesion (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of cohesion (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Significance of the Study

The major principle that establishes the foundation of the present study is that good lexis is a critical factor that influences the holistic quality of a piece of writing (Laufer, 2013). By focusing on the lexical features in native and nonnative speakers' writing samples, the current study is significant for 1) better understanding the lexical characteristics of learner language in academic English writing; 2) identifying the differences between NNSs and their native speaking peers in academic writing; 3) detecting the potential influences from L1 or educational backgrounds on academic English writing; 4) and developing targeted instructional strategies for academic writing according to learner needs.

First, sufficient corpus-based data provide a solid basis for analyzing naturally-occurred languages. The corpora employed in the present study include 1) the International Corpus of Learner English (ICLE), which represents the academic writing of NNEs; 2) the Louvain Corpus of Native Essay Writing (LOCNESS), which represents the academic writing of NESs. By using advanced computational techniques, the current study examines three lexical features (i.e., lexical diversity, lexical sophistication, and cohesion) of both native and nonnative speakers' writing products. Hence, a general description of lexical quality of both population groups can be revealed.

Second, with the rich description of lexical features, comparisons are conducted to illustrate the differences between NESs and NNEs in terms of lexical use in academic writing. Although native English writing does not equal flawless academic writing by any means, acknowledging the differences between native and nonnative writers is informative for both ELs and instructors to establish a realistic goal of achieving higher writing performance strategically. Moreover, to recognize the diversity of ELs, comparisons are also made across all writers from different language backgrounds. The potential differences may provide insights in addressing influences from their mother tongues or educational backgrounds, thus further inform the corresponding pedagogies.

Lastly, the deep and thorough analyses reveal the characteristics of learners' written products and their needs in improving lexis, this also contains crucial pedagogical implications. Based on learner characteristics and their needs, more targeted and effective instructional strategies can be developed.

Study Limitations

Several limitations apply to the present study. First, the researcher (I) did not participate in the compiling of the two corpora used in the study. Both corpora (i.e., the ICLE and the LOCNESS) were compiled under the supervision of Sylviane Granger at the Université Catholique de Louvain. The published descriptions of both corpora are precise, and they meet the research purpose of the current study. However, the lack of personal participation could cause misunderstanding and misinterpretation of the corpora. Detailed descriptions of both corpora are provided in the subsequent chapter, allowing readers to have a better understanding of these corpora and evaluate the reliability of employing the corpora in the current study.

Second, owing to the scope and specific focus of the present study, only argumentative writing samples are selected from the ICLE and the LOCNESS to represent academic writing. Currently, argumentative writing is a major form of writing for most ELs who intend to enroll in English-medium higher educational institutions. Even though argumentative writing is common across all disciplines in academic contexts, academic writing includes other genres. For instance, requirements and writing characteristics of informative and narrative essays may differ from argumentative essays. Thus, the present study does not cover the whole scheme of academic writing. For future research, it is encouraged to explore lexical characteristics in other academic writing genres.

Lastly, due to the large number of essays included in the study (700 essays in total), it is extremely difficult to manually check the spelling mistakes in all writing samples. Thus, the spelling errors might slightly skew the measures of the lexical features in some cases. Also, there are a few spelling differences between British and American university students' essays. The

present study ignores these differences and spelling errors because they do not account for a major difference of the final results.

Definitions of Terms for the Study

The following terms and acronyms appear frequently throughout the dissertation. For better understanding of the study, the definitions of the terms are provided below.

- *Academic Vocabulary List (AVL)*: a word list developed by Gardner and Davies (2013), which contains academic vocabulary lists of English that are based on 120 million words of academic texts in the Corpus of Contemporary American English (COCA).
- *Academic Word List (AWL)*: a word list developed by Coxhead (2000), which includes 570 word families and is considered to have a high coverage in all academic prose.
- *English for Academic Purposes (EAP)*: English language instruction for academic study.
- *English as a Foreign Language (EFL)*: EFL learners refer to the learners in a country where English is not the dominant nor native language.
- *English as a Second Language (ESL)*: ESL learners refer to the learners in a country where English is the dominant or native language.
- *English Learners (ELs)*: learners who study English as a second or foreign language. It is interchangeable with English Language Learners (ELLs), Nonnative Speakers (NNSs), and Nonnative English Speakers (NNEs) in this dissertation.
- *First language (L1)*: the native language of the learner acquired from birth; mother tongue
- *Intensive English Program (IEP)*: language learning centers that prepare learners for postsecondary study in English in a university where English is the native language.
- *Learner Corpus Research (LCR)*: research carried out based on learner corpora

- *Lemma*: a group of word forms that are related by being inflectional forms of the same base word. For instance, the verb *destroy* can be considered as a base word; its inflected forms, including *destroys*, *destroying*, and *destroyed*, are all part of the verb lemma. However, the word *destruction* is considered as a separate lemma.
- *Lexical Frequency Profiles (LFP)*: a tool developed by Laufer and Nation (1995) to measure lexical richness (i.e., lexical sophistication) in learner writing.
- *Native speakers (NSs)/ native English speakers (NESs)*: NSs of English refer to speakers who are either monolingual and/or speak English as the first language.
- *Nonnative speakers (NNSs)/ nonnative English speakers (NNEs)*: NNSs of English refer to speakers who speak English as the second or foreign language.
- *Second language (L2) / foreign language (FL)*: the additional language learned some time after the learner's first language
- *Teaching English to Speakers of Other Languages (TESOL)*
- *Tokens/running words*: the total number of words in a text
- *Types*: the total number of different words in a text. A type is also called an individual word form. For instance, *help* and *helps* are two different types.
- *Word family*: different from the concept of lemma, the concept of word family includes both inflectional and derivational word forms. For instance, *destroy*, *destroys*, *destroying*, *destroyed*, and *destruction* are considered as one word family.

Organization of the Study

Chapter One introduces the background and rationale of the current study. The background information and rationale establish the foundation of the research questions. The major research questions and corresponding hypotheses are also presented in this chapter. Next,

the chapter presents the significance of the study from pedagogical perspectives and the limitations of it. Finally, the chapter includes definitions of key terms in the study and the overall organization of the dissertation.

Chapter Two reviews the scholarship and prior literature related to the research area in the present study. This chapter first discusses the uniqueness of academic language and the critical role that vocabulary plays in achieving higher performance in academic skills. Next, theoretical and methodological foundation of the present study is addressed, which leads to the necessity of conducting corpus-based study for the research questions. In addition, the research questions require related reviews of learner corpora research and the corresponding research method, Contrastive Interlanguage Analysis (CIA). Based on the research objective of presenting a comprehensive picture of lexical features in academic writing, the classification and detailed explanations of lexical features in the literature are introduced. Lastly, corresponding measurements of each targeted lexical measure in the current study are presented.

Research design and selected corpora are introduced in Chapter Three. This chapter also includes sampling method and procedures of data collection. The next major section of this chapter details the instruments that are used to measure the lexical features of the academic writing samples.

The results of the statistical analyses are presented in Chapter Four. Chapter Five serves as the conclusion of the dissertation. It concludes and summarizes the study by further discussing the findings and limitations of it. Pedagogical implications and recommendations for future research are lastly provided to close the dissertation.

CHAPTER TWO: RESEARCH AND LITERATURE REVIEW

The past few decades have witnessed a dramatic increase in the number of ELLs across different educational levels. In the 2014 – 2015 school year, for example, an estimated 4.6 million students were ELs (National Center for Education Statistics, 2018), representing 9.4% of public school students in the United States. This increase is not an anomaly. Ten years ago, this percentage was 9.1%, or an estimated 4.3 million students; thus, the increase is relatively steady. From the year 2014 to 2015, some states (e.g., California, Nevada, and Texas) experienced an even larger share of EL population with percentages reaching 10% or more (National Center for Education Statistics, 2018). At the post-secondary level, from 2015 to 2016, 5% (estimated to be more than one million) of the more than 20 million students enrolled in U.S. higher educational institutions were international students coming from non-English language background (Witherell & Department of State, 2016). Compared to just a decade ago, the percentage of international students studying at U.S. higher educational institutions has increased 85%.

Beyond the United States, 1.75 billion people, a quarter of the world's population, speak English at a useful level (British Council, 2013). Thus, there is no doubt that English has become today's global lingua franca. English is used as a crucial language of communication across various settings; being able to use English in an academic setting is an inevitable task for many NNEs worldwide. Compared to the year 2000, the population of students traveling to other countries for tertiary-level studies doubled in 2012, reaching four million in total (UNESCO Institute for Statistics, 2016). Not surprisingly, English-speaking countries are the most popular destinations for international students (Pop, 2016). To evaluate foreign students' English language proficiency and ensure their academic performance after admission, most universities require students to provide some sort of standardized English proficiency evaluation to prove

their academic English proficiency (Kice, 2014). Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) are two commonly accepted examinations. Hence, NNEs encounter the challenge of improving their academic English proficiency in order to secure admission from the universities abroad.

Besides mandatory requirements from English-speaking countries for NNEs, the internationalization of institutions in higher education encourages universities in non-English-speaking countries to provide English-medium courses. This market-driven move is believed to be beneficial for attracting foreign students, enhancing the university's international profile and prestige, as well as improving English language skills of domestic staff and students (Ferguson, 2007). Other than acquiring academic English for passing gatekeeping examinations, being able to use English for communicative purposes in academia is also an essential construct in academic settings. In the world's largest academic journal indexing system, *Web of Science*, it is noted that publishing full text in English is the apparent trend for international research community (Testa, 2016). In sum, in the U.S., other English-speaking countries, and non-English-speaking countries, the necessity of mastering English for academic purposes is widely recognized. Hence, the present study focuses on investigating issues of English learning and the use of English in an academic context.

This chapter first reviews the nature of academic language and the central position of vocabulary in the development of L2 proficiency. The relationship between vocabulary and the four skills of language learning, namely listening, reading, speaking, and writing, is discussed to demonstrate the critical role of vocabulary. In particular, extensive empirical support will be used to demonstrate the centrality of vocabulary in L2 writing development, which is at the heart of the current study. Next, a review of the scholarship on the Contrastive Interlanguage Analysis

(CIA) (Granger, 1996) and related corpus linguistics studies is presented to set the theoretical and methodological foundation of the present study. The last major section of this chapter focuses on addressing the measures of lexical features in academic writing and introducing three target lexical features of the present study.

The Nature of Academic Language

As an interdisciplinary construct, academic language has evolved from various academic fields. Linguistics, Applied Linguistics, and Education are three major disciplines where academic language is widely explored. Even though these fields have differentiated emphases and one single definition of academic language has yet to emerge, researchers have agreed on the foundational nature of academic language across disciplines.

In the field of SLA, one of the most cited pieces of scholarship to date is Cummins' (1979) dichotomy of BICS and CALP. Basic Interpersonal Communication Skills (BICS) refers to informal daily language usage, which is naturally acquired within the L1 (Cummins, 1979). For L2 learners, BICS is relatively easy and usually the first to be acquired (Ellis, 2008). In contrast, Cognitive Academic Language Proficiency (CALP) emphasizes the formal usage of language, which requires a much longer period of time to reach a desired proficiency level.

The features of formal discourse found in CALP are mostly revealed via sophisticated vocabulary and grammatical structures (Coxhead, 2000; Folse, 2004; Nation, 2001; Scarcella, 2002). Bailey (2007) characterizes academic proficiency as closely related to the ability to use both general and specialized vocabulary, grammatical structures, and discourse structures. After synthesizing a wide range of literature on academic language, Anstrom et al. (2010) conclude academic language as “development, with trajectories of increased sophistication in language use

from grade to grade, with specific linguistic details that can be the same or vary across content domains” (p. 4).

Overall, academic language is a complex system that consists of various linguistic, cognitive, and metacognitive skills (Singhal, 2004). Among these diverse components, the linguistic dimension is considered as the foundational construct in order to achieve proficient use of academic language across disciplines. Furthermore, within the linguistic component, there are three major aspects that determine the integral competency of academic language, namely lexical, grammatical, and discourse features (Bailey & Butler, 2003; Bailey, Butler, & Sato, 2005). In the present study, lexical features in academic language are focused due to the broadly acknowledged critical status of vocabulary in language learning. Next, the central position of vocabulary in developing the four language skills (i.e., listening, reading, speaking, and writing) is presented from the perspective of learning English as a second language.

The Role of Vocabulary in SLA and TESOL

From being traditionally neglected in the area of SLA to becoming one of the central topics discussed by researchers and practitioners (Meara, 2002), research in vocabulary teaching and acquisition has risen to a new level (Nation & Webb, 2010). In the last 120 years, over 30% of L1 and L2 vocabulary research occurred in the last 12 years (Nation, 2011). With research conducted through the assistance of computational linguistic tools, the fundamental role of vocabulary in SLA has been well established.

In this section, the role of vocabulary in learning English as a second language is presented from the development and performance of four language skills, namely listening, reading, speaking, and writing. Vocabulary in developing listening, reading, and speaking skills

is discussed briefly in one section. Vocabulary in developing writing skills is then reviewed extensively.

Vocabulary and Listening, Reading, and Speaking

Empirical studies in the past two decades have employed various testing instruments, methodology, research foci, and contexts to investigate the correlation between vocabulary knowledge and listening comprehension competence. The studies consistently share common findings, confirming the strong positive correlation between vocabulary knowledge and listening comprehension; thereby supporting the critical position of vocabulary in listening. Five selected empirical studies from the recent body of literature are presented in Table 1 to elucidate the role of vocabulary in listening comprehension.

Table 1

Vocabulary and Listening

Study	Main Findings
Bonk, 2000	- Participants ($N = 59$) failed to achieve high comprehension scores when fewer than 75% of the lexical words in the input texts were recognized.
Matthews & Cheng, 2015	- Significant strong and positive correlation between Word recognition from speech (WRS) scores at each frequency level and listening scores was observed ($r = .67$ (K1), $.69$ (K2), $.72$ (K3), $p < .01$). - K3 WRS scores were able to predict more than half (52%) of the variance in listening scores ($F(1,165) = 180.90$, $p < .001$, $R^2 = .52$). - High-frequency vocabulary plays a fundamental role in listening comprehension
Milton, Wade, & Hopkins, 2010	- Correctly answering questions on the listening test required learners to master vocabulary knowledge in decoding both written and auditory content.
Stæhr, 2009	- Vocabulary size and depth of vocabulary knowledge were both significantly correlated with listening comprehension ($N = 115$; $r = .70$ and 0.65 ; $p < .01$). - Vocabulary size and depth of vocabulary knowledge together accounted for over half (51%) of the variance in the listening scores.
Wang & Treffers-Daller, 2017	- Detected the strongest correlation between vocabulary size and listening comprehension ($r = .44$). - The robust predicting effect of vocabulary knowledge in listening comprehension was supported.

The crucial role of vocabulary knowledge in L2 reading comprehension has also been well acknowledged by researchers in empirical studies. The bottom line is that the more vocabulary that is known to a reader, the less difficulty the reader may encounter during reading in the target language. Several relatively influential and recent studies in the field are selected to address the mainstream stance of vocabulary in developing L2 reading comprehension. The studies are summarized in Table 2.

Table 2

Vocabulary and Reading

Study	Main Findings
Laufer, 1992	<ul style="list-style-type: none"> - Highly significant correlation existed between reading and vocabulary scores ($N = 92$, $r = .5$ and $.75$, $p < .0001$). - 3,000 word families should be considered as the turning point of vocabulary size for reading comprehension as it predicted more than half of the reading scores (56%).
Nassaji, 2004; Leider et al.'s, 2013	<ul style="list-style-type: none"> - Learners who had stronger depth of vocabulary knowledge were able to use certain lexical inferencing strategies more frequently. - The depth of vocabulary knowledge contributed to effective inferencing process in L2 reading.
Qian, 2002	<ul style="list-style-type: none"> - Vocabulary depth knowledge was as important as vocabulary size in predicting reading comprehension scores, with both factors accounting for more than 50% of the variance.
Zhang & Anual, 2008	<ul style="list-style-type: none"> - Significant correlation existed between vocabulary knowledge of high-frequency words (2,000-word level) and reading comprehension ($r = .423$, $p < .01$). - At 3,000-word level, highly significant correlation was found for the performance of short-answer questions ($r = .848$, $p < .01$).

As in the listening and reading components of L2 development, vocabulary knowledge plays a key role in impacting the development of L2 learners' speaking proficiency as well. Effective oral fluency is largely built upon sufficient vocabulary size and depth. Empirical research has been conducted to demonstrate the relationship between vocabulary knowledge and L2 speaking performance. A few related key studies are introduced in Table 3 to reveal the critical role of vocabulary in L2 speaking.

Table 3

Vocabulary and Speaking

Study	Main Findings
Hilton, 2008	<ul style="list-style-type: none"> - The speakers' vocabulary knowledge had a significant positive correlation with their speech rate, which was measured by words spoken per minute ($N = 56$, $r = .581$, $p < .0001$). - The more words a speaker knows, the more fluently he/she can speak.
Koizumi & In'nami, 2013	<ul style="list-style-type: none"> - The structural equation modeling confirmed the substantial role of vocabulary breadth and depth in explaining the variance in speaking proficiency. - The two studies involved in this research showed respectively 32% and 64% effect size in predicting the variance of speech fluency through vocabulary knowledge.
Lourdunathan & Menon, 2017	<ul style="list-style-type: none"> - The participants' limited vocabulary knowledge hindered the effectiveness of interactive strategies. - The researchers recommended the teaching of essential vocabulary and the employment of interactive strategies.
Yu, 2009	<ul style="list-style-type: none"> - Lexical diversity was significantly related to the overall scores of interviews, namely speaking performance in this study ($r = .484$, $p < .01$).

Vocabulary and Writing

At the higher education level, it is almost impossible to avoid writing. Students might be able to avoid speaking in the class or communicating orally with their professors. However, almost all courses require some types of written assignments to prove the student's understanding of the content. Therefore, writing well in the target language is a primary task that L2 learners need to achieve in higher education (Coxhead & Byrd, 2007). Based on this rationale, the present study focuses on L2 writing and explores the strategies to develop NNESS' writing skills. This section aims to synthesize and provide a deeper review of empirical scholarship on how vocabulary strongly influences L2 writing performance (see Table 4 for summary).

The process of progressing from the initial encounter of a word to being able to use it in writing is multidimensional. It involves the ability of understanding and expressing the word in a

range of contexts as well as using correct spelling and grammatical collocations of the word (Coxhead & Byrd, 2007). In addition, Hyland (2007) discusses another facet of the lexical puzzle – genre – as he emphasizes the importance of successfully distinguishing language and lexical features in various genres and academic disciplines. On one hand, scholarship in examining the relationship between vocabulary and L2 writing has obtained findings in concert with the other three language skills, concluding the vital role of vocabulary in developing L2 writing skills (Laufer, 2013). On the other hand, more research in L2 writing offers empirical support regarding particular aspects of vocabulary that need to be developed and the corresponding pedagogical strategies.

One of the pioneer studies on vocabulary size, lexical richness, and L2 writing quality was conducted by Laufer and Nation (1995). This study was based on word frequency in English. Laufer and Nation developed and tested their Lexical Frequency Profiles (LFP), which was a system to evaluate the differences of lexical quality in L2 writing of learners from various proficiency levels ($N = 65$). The results suggested that compared to higher-proficiency level participants, participants in the low-proficiency group used a significantly larger number of first 1,000 words, and fewer numbers and types of academic words and off-list words (all $p < .05$). The researchers proposed that it is reasonable to expect that the L2 writers' writing products reflect their vocabulary size.

Stæhr (2008) also focused on the relationship between vocabulary size and the skills of listening, reading, and writing. In terms of writing, each participant in Stæhr's study ($N = 88$) wrote a 450-word composition as a measure of their writing skills. The correlation between vocabulary size and the writing scores was found to be strong and significant with a coefficient of .73. Meanwhile, 52% of the variance in receiving an average or above-average writing score

was explained by vocabulary size. Thus, the results confirmed the critical role of receptive vocabulary size in enhancing writing performance.

Besides receptive vocabulary size, the correct way of using certain vocabulary in writing is another important aspect of truly knowing a word. Engber (1995) explored how lexical richness of an essay influenced the holistic quality of the writing from the readers' perspective. Sixty-six timed essays from an IEP placement test were collected to be evaluated by trained IEP teachers from a holistic perspective. The lexical richness analyzed in the essays included lexical variation, lexical error, and lexical density. The results suggested the significant effect of lexical variation and correctness on readers' judgement of the essays. Error-free variation showed the highest correlation with final scores of the essays ($r = .57, p < .01$). This indicates that not only the vocabulary size matters in improving L2 writing quality, but also the correct use of vocabulary.

Zhou (2009) took the ESL learners' perspective to examine their perceptions and goals in improving academic writing ($N = 15$). Through semi-structured interviews and stimulated recall sessions, students in the Canadian pre-university EAP program unanimously expressed their belief in the importance of using appropriate vocabulary in writing. The participants stated that having the knowledge of academic words can be beneficial for them to precisely express their ideas in writing. This perception from L2 learners' perspective is supported by Coxhead (2012), which also revealed that learners were all aware of the key role of using academic and professional vocabulary to express their ideas in writing appropriately; however, the techniques that the participants were able to employ to incorporate academic vocabulary in their writing were rather limited.

The confirmed importance of vocabulary for L2 writing has led researchers to explore specific types of vocabulary knowledge that are needed for improving writing quality. Johnson, Acevedo, and Mercado (2016) suggest that accurate productive knowledge of high-frequency vocabulary correlates with L2 writing performance, while the use of high-frequency vocabulary indicates less developed writing quality.

Another unique way to examine the demanding vocabulary for writing is from the perspective of writing genres. Different uses and selections of vocabulary are considered as a distinguishing feature that represents various genres of writing (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Olinghouse and Wilson (2013) explored the relationship between various vocabulary constructs (i.e., diversity, maturity, elaboration, academic words, content words, and registers) and three genres of writing (i.e., story, persuasive, and informative writing). First, the comparison on the measurements of vocabulary across genres did show differences. For instance, persuasive texts contained higher diversity than informative texts. Second, analysis on measures of vocabulary that predict writing quality suggested: 1) vocabulary diversity was a significant predictor for story writing quality; 2) content words and registers were significant predictors for persuasive writing quality; 3) content words were the strongest predictor for informative texts.

In sum, for pedagogical implication, when emphasizing the importance of vocabulary for writing quality, specific writing genres should be distinguished to address the corresponding vocabulary needs for different genres.

Finally, research on theoretical relationship between vocabulary and writing promotes further studies with pedagogical purposes. Muncie (2002) employed the LFP from Laufer and Nation's (1995) study to investigate whether L2 learners' writing quality could be enhanced by

improving their vocabulary knowledge. The study examined three drafts of the participants' timed compositions ($N = 25$). The results revealed a significant drop of below 1,000 level words and an increase of above 2,000 level words ($p < .05$), which showed that the revision process indeed helped the participants use a larger proportion of sophisticated words and further improve the holistic quality of their writing products. The pedagogical implication from this study is that during the pre-writing stage, vocabulary instruction and preparation is necessary and essential.

In another empirical study, Webb (2009) supported the positive effect of pre-learning vocabulary on writing. The participants of the study were 71 EFL learners in a Japanese university. Supported by the studies of Lee (2003) and Snellings, van Gelderen, and de Glopper (2004), as well as research on L1 writing by Yonek (2008), the results of Webb's study demonstrated that the participants were able to correctly use 35% of the target words that were taught in the productive word learning session, which was higher than the receptive learning group. Therefore, Webb's study further indicates the effectiveness of vocabulary instruction before writing. Table 4 concludes the empirical studies that have been conducted to support the importance of vocabulary in academic writing.

Table 4

Vocabulary and Writing

Study	Participants	Research Questions	Method	Main Findings
Engber, 1995	- 66 essays written by intermediate to high-intermediate IEP students from various language backgrounds.	- The role of the lexical component as one factor in holistic scoring.	- Raters graded the essays for holistic quality of the writing samples. - Lexical richness measures: lexical density, error-free variation, percentage of lexical error, lexical variation,	- The results suggested the significant effect of lexical variation and correctness on readers' judgement of the essays. - Error-free variation showed the highest correlation with final scores of the essays ($r = .57, p < .01$).
Laufer & Nation, 1995	- ELs in New Zealand ($n = 22$) and Israel ($n = 43$). - Learners in NZ were from various language backgrounds. - Participants were in various proficiency levels.	- Will there be a significant difference between the LFP of different language proficiency levels? - Will the LFP of the compositions correlate highly with the scores of the same learners on the active version of the VLT?	- Each participant wrote 2 compositions; took the Vocabulary Levels Test (VLT). - Analyzing the compositions by the LFP.	- Participants in the low-proficiency group used a significantly larger number of first 1,000 words, fewer numbers and types of academic words and off-list words (all $p < .05$). - L2 writers' writing products reflected their vocabulary size.
Stæhr, 2008	- 88 EFL learners from lower secondary education in Denmark	- To what extent is vocabulary size associated with the skills of listening, reading and writing? - Is it possible to determine a vocabulary size	- Each participant wrote a 450-word composition. - 2 weeks prior to the examination, participants completed a vocabulary size test (the VLT).	- The writing scores were significantly and highly correlated with vocabulary size, producing a coefficient of .73. - The result suggested a relatively strong relationship between learners' vocabulary

Study	Participants	Research Questions	Method	Main Findings
		threshold above which learners are likely to perform above average in the reading, listening and writing test?		size and the quality of their written compositions.
Zhou, 2009	- 15 EAP students in Canada.	- What types of vocabulary do learners want to improve in EAP and university courses? What actions do learners take to improve their vocabulary?	- Semi-structured interview and stimulated recall sessions conducted at the beginning and end of a writing course.	<ul style="list-style-type: none"> - Knowing the first 2000 words in English made a difference in writing scores. - The participants unanimously expressed their belief of the importance of using appropriate vocabulary in writing. - The participants expressed their demands of mastering more academic words.
Others:				
- Coxhead, 2012				
Learners were all aware of the key role of using academic and professional vocabulary to express their ideas in writing appropriately.				
The participants had limited ability of incorporating academic vocabulary to their writing.				
- Johnson, Acevedo, & Mercado, 2016				
Accurate productive knowledge of high-frequency vocabulary correlated with L2 writing performance.				
The use of high-frequency vocabulary indicated less developed writing quality.				
- Lee, 2003; Snellings et al., 2004; Webb, 2009; Yonek, 2008				
Vocabulary instruction before writing had positive effects on writing quality.				
- Muncie, 2002				
During the pre-writing stage, target vocabulary instruction and preparation was necessary and essential.				
- Olinghouse & Wilson, 2013:				
Persuasive texts contained higher diversity of vocabulary than informative texts.				
Vocabulary diversity was a significant predictor for story writing quality.				
Content words and registers were significant predictors for persuasive writing quality.				
Content words were the strongest predictor for informative texts.				

Summary

This detailed review of scholarship clearly indicates the critical contribution of vocabulary to the development of the four skills of L2, including listening, reading, speaking, and writing, but especially for writing. For various populations in different contexts, mastering sufficient size and depth of vocabulary is essential to achieve high proficiency of all aspects of the target language. The empirical findings of vocabulary in L2 development, especially in writing, establish the rationale of the present study in terms of focusing on lexical aspects of native and nonnative writing.

Next, a review of the scholarship on the Contrastive Interlanguage Analysis (CIA; Granger, 1996) and related studies of Corpus Linguistics are presented as the theoretical and methodological foundation for conducting the current study.

Theoretical and Methodological Foundation of the Present Study

This section begins with a general introduction of the field of Corpus Linguistics and its development in SLA, especially in TESOL. Following the existing corpus studies of learner English, the theoretical framework of the Contrastive Interlanguage Analysis (CIA) and its implementation are discussed. Last, the connection between the present study and the CIA framework is revealed to lead to the research foci of the present study.

Corpus Linguistics, SLA, and TESOL

Biber and Reppen (2015) describe Corpus Linguistics as a research approach to explore language variation and use empirically. Thanks to the large and principled collection of natural texts as well as computational quantitative and qualitative analysis techniques, the findings based on corpus linguistics approach are generalizable, reliable, and valid. Starting from the 1980s, the increasing of computational tools facilitated the development of large electronic corpora and

systematic analysis of these corpora became possible. Hence, major linguistic studies began to appear in the 1980s (Biber & Reppen, 2015). In this section, I present the major research projects in Corpus Linguistics and the application of corpus-based approach in SLA and TESOL.

Major corpora and research in Corpus Linguistics

Before the 1980s, in a relatively unsupportive environment in the field of linguistics, Quirk (1960) stated the necessity and benefits to construct the Survey of English Use (SEU) as a descriptive and systematic reference for English users to follow other than their uncertain intuition. This is considered as the start of English Corpus Linguistics. Also, starting in the early 1960s, Francis and Kučera began working on the Standard Corpus of Present-Day American English, later known as the Brown Corpus. The compiling took almost two decades to complete. By 1979, the corpus consisted of 1,014,312 running words from English prose printed in the United States during the year 1961 (Francis & Kučera, 1979). Even though the construction of these early corpus works may not seem substantial when considering the possibility that a computer can function as an aggregator nowadays, they most certainly played a pioneering role in inspiring the following important research in Corpus Linguistics.

Employing computational tools, the British National Corpus (BNC) was built during 1991 to 1994, containing 100-million-word collection of samples of written and spoken English language. The samples are from a wide array of sources to represent British English from the later part of the 20th century. Another recent important corpus is the Corpus of Contemporary American English (COCA, Davies, 2008-). By 2017, the COCA contained more than 560 million words of texts. It is equally divided among various registers, including popular magazines, newspapers, and texts from spoken, fiction, and academic works. By far, tremendous types of corpus for various purposes have been compiled. The development of these corpora

offers a solid foundation for research in Corpus Linguistics to further explore language use and variation from an empirical perspective.

One important application of Corpus Linguistics is the research in grammatical patterns of language. Biber et al.'s (1999) *Longman Grammar of Spoken and Written English (LGSWE)* is the quintessential grammar reference book in the field of English language studies. This book contains references based on computer-aided and corpus-based information. The corpus used for *LGSWE* is the Longman Spoken and Written English Corpus, which contains over 40 million words of texts representing six major registers. By examining the use of grammatical features in both American and British English, this book has become one of the most fundamental products of corpus-based research due to its inclusive analyses of both spoken and written texts across various register categories. Following the same approach, *Cambridge Grammar of English* (Carter & McCarthy, 2006) also applied corpus-based analyses to demonstrate how grammatical features can be described across spoken and written registers.

Another remarkable area of utilizing corpus approach extensively is lexical studies. Many earlier applications of corpora were to provide word lists that represent frequency features of the language (Francis & Kučera, 1982; Johansson & Hofland, 1989). With help from computational techniques, compiling word lists for various purposes has become fairly manageable. One of the most popular word lists today is the Academic Word List (AWL; Coxhead, 2000). This list includes 570 word families and is considered to have a high coverage in all academic proses. Later, researchers also developed discipline-based academic word lists for special purposes (Chen & Ge, 2007; Li & Qian, 2010). Other examples of corpus-based vocabulary study include the New General Service List (Browne, Culligan, & Phillips, 2013), the Business Word List

(Konstantakis, 2007), the Basic Engineering List (Ward, 2009), and the Phrasal Expressions List, or PHRASE (Martinez & Schmitt, 2012).

In addition to surface level studies based on word frequency, corpus-based investigation on collocation has provided empirical insights in understanding word meanings and usages, such as Sinclair (1996) and Partington (1998). Further studies also considered register differences while analyzing word collocational associations. For instance, Biber, Conrad, and Reppen (1998), Gledhill (2000), and Marco (2000) all discussed the functions of collocations in academic research writing. In short, with deeper and wider applications of corpus approach, language studies have been enhanced drastically.

Language changes are also studied more deeply with the growing corpora and computational techniques. Leech, Hundt, Mair, and Smith (2009) systematically studied English language change over a precisely defined period of time in the recent past (early 1960s to early 1990s). Corpus-based historical research approach provides the study with rigorous methodology to detect how English grammar has changed over time. The corpora data that helped the research reach a conclusive statement include four matching corpora from the “Brown Family”: The Brown corpus, the Lancaster-Olds/Bergen corpus (LOB), the Freiburg-Brown corpus (FROWN), and the Freiburg-Lancaster-Olds/Bergen corpus (FLOB). Due to the nature of these corpora and their parallel characteristics, the researchers were able to track the changes in both American and British English, as well as the connections between these two types of English. Other corpora (e.g., the Diachronic Corpus of Present-Day Spoken English) were also employed to detect the changes in spoken English (Leech et al., 2009).

In terms of sociolinguistics, even though a traditional qualitative approach is still widely used in most studies, a few studies on the change of regional dialects have employed a corpus

approach as well (Biber, Reppen, & Friginal, 2010). For instance, the Newcastle Electronic Corpus of Tyneside English (NECTE; Corrigan & Buchstaller, 2007), the Helsinki Corpus of English Texts (Rissanen, 1993), and the Freiburg English Dialect Corpus (FRED, Anderwald & Wagner, 2007; Kortmann & Wagner, 2005) were compiled to provide corpus data for research in sociolinguistics. Furthermore, research on global varieties of English has been carried out mostly from a corpus-based perspective. The International Corpus of English (ICE; Greenbaum, 1988-) is considered as a representative research project of World Englishes. The research is an ongoing project, which intends to compile parallel corpora for all varieties of English around the globe.

In sum, the supportive empirical research environment and advanced computational techniques bolstered the expansion of the field of Corpus Linguistics in both width and depth. Meanwhile, the nature of the corpus approach enables researchers to explore issues of linguistics reliably and validly. For the purpose of the present study, next section specially focuses on reviewing major influential SLA and TESOL studies that are based on corpus data.

Corpus-based research in SLA and TESOL

Along with the advancements in corpus-based technology and research in linguistics, explorations into pedagogical applications of Corpus Linguistics continue to grow. From a language teaching perspective, corpus-based language teaching materials for various language skills have been created and accepted in mainstream language classrooms as well as language learning for specific purposes. From a language learning perspective, data-driven learning (DDL) based on corpus data has been earning focus and discussion among language educators and learners (O’Keeffe & McCarthy, 2010). Moreover, the creation of learner corpora has received much attention in empirical research that advocates its pedagogical value. In this section, I focus

on synthesizing the major development of corpus-based research and applications in SLA and TESOL.

Leech (1997) notes the value of converging language corpora and teaching. There are three foci of the convergence that Leech suggested: 1) the indirect use of corpora in teaching, including reference publishing, syllabus designing, material developing, and language testing; 2) the direct use of corpora in teaching, such as using corpora for hands-on classroom activities and encouraging students' individual interaction with corpora; 3) teaching language for domain-specific usages and professional communication.

In the field of TESOL, a few corpus-based textbook series have been widely accepted in both EFL and ESL contexts. *Touchstone* series (McCarthy, McCarten, & Sandiford, 2005) is a successful English textbook series that is entirely based on corpus evidence. It demonstrates how everything from syllabi to practical exercises can be designed based upon corpus research. *Grammar and Beyond* series written by Reppen and Gordon (2011) is another popular grammar textbook series for ESL institutions. This series includes corpus-based information and language usages that can be found in real-world contexts.

A dictionary is another necessary tool for L2 learning. Starting from the early 1990s, learner dictionaries have become popular among language learners and educators due to their appropriate comprehension level and representation of naturally-occurred language. *Cambridge Learner's Dictionary* (Woodford, 2001) and *Longman Dictionary* series are two remarkable examples of learner dictionaries.

Language testing is also a crucial component of L2 teaching and learning in academic contexts. The establishment of large and comprehensive corpora have been employed as an archive of language examination scripts. The application of corpus data is beneficial for

optimizing test procedures, improving the quality of test marking, validating and standardizing tests (McEnery & Xiao, 2011).

In terms of direct use of corpora in L2 teaching, corpus interfaces (e.g., COCA) and programs (e.g., Compleat Lexical Tutor) are designed to provide language classrooms with hands-on activities that stimulate learners' engagement in the language discovery and learning process (John, 1991). The major advantages of introducing corpus tools to language learners include improving active and autonomous learning as well as offering explicit instruction of language patterns and rules based on authentic concordance information (John, 1991; Kennedy, 2001).

Lastly, the fields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP) have also been extensively enhanced thanks to the advancements of Corpus Linguistics. For instance, Upton and Connor (2001) carried out a "move analysis" in the field of Business English based on a business learner corpus. The study examined and compared the politeness strategies that were used by ELs from various cultural backgrounds. Thus, the study achieved the sociolinguistic component of the cultural aspect of communication and language usage in professional fields.

Another key application of Corpus Linguistics in SLA and TESOL is to connect with learners' perceptions and needs. Data-driven learning (DDL) (John, 1991) was developed as an approach to motivate learners to become active learners and even language researchers. Introducing corpus tools to language learners has been empirically proved to be an effective DDL approach (St John, 2001). DDL has been confirmed with the potential to increase learners' autonomy in language learning and further improve their language proficiency.

Last but not least, the relatively recent creation of learner corpora directly connects Corpus Linguistics to L2 language learning and teaching. The collections of spoken and written learner language offer a solid basis for systematic analyses of interlanguage development, which moves the focus of corpus-based language studies from native speaker dominance to the language performance of learners (O’Keefe & McCarthy, 2010). The purpose and research questions of the present study call for the involvement of learner and native speaker corpora to examine what ELs may lack in terms of lexis in academic writing.

In sum, the development of Corpus Linguistics drives its application in SLA and TESOL. The extensive employment of corpus-based approaches provides effective and rigorous resources for both instructed and self-guided language learning. The value has been recognized by language researchers, practitioners, and learners. In the next section, related research on learner corpora will be presented in further detail, and the Contrastive Interlanguage Analysis (CIA; Granger, 1996) will be presented as a research framework for analyzing learner corpora.

Learner Corpus Research (LCR) and Contrastive Interlanguage Analysis (CIA)

Over the past three decades, the compilation of diverse corpora and the progression of computational techniques have driven the research of corpus-based language studies considerably. However, a majority of the corpora used in the empirical studies were native speakers’ written or spoken texts; the nonnative language varieties had been largely neglected until the late 1980s (Granger, Gilquin, & Meunier, 2015). As an offshoot of Corpus Linguistics, research on learner corpus has risen and shed light on L2 pedagogy by providing large and naturally-occurred data of learner language. Led by Sylviane Granger and her colleagues since the late 1980s, the field of learner corpus research (LCR) has been acknowledged with substantial values and continues to grow steadily.

This section reviews the development of LCR and its major methodological framework of analysis, Contrastive Interlanguage Analysis (CIA). The conclusion is made by connecting the rationale and theoretical foundation of the present study to the employment of CIA.

Learner Corpus Research (LCR): Major Concepts and Studies

Coming from the purpose to develop more learner-aware and learner-focused language teaching and learning materials, the emergence of learner corpora, especially computer learner corpora, began two decades ago (Granger, Gilquin, & Meunier, 2015). Building upon Sinclair's (1996) definition of corpora, Granger (2002) defined computer learner corpora as "electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose". Having the definition and requirements of learner corpora established, research carried out based on learner corpora is grouped under the umbrella term of Learner Corpus Research (LCR; Granger, Gilquin, & Meunier, 2015). Using the techniques of Corpus Linguistics, LCR focuses on describing and examining learners' linguistic performances.

One major advantage of LCR is the potential to analyze large collections of authentic learner language productions, which results in better representation of the population performance and provides a firm basis for various analyses. In addition to presenting what the learners are able to produce, another unique contribution of LCR is to show what language aspects learners fail to produce and what kind of errors often occur (Cobb & Horst, 2015). In a nutshell, a learner corpus has the potential to provide solid data that can reveal learner needs and thereby further ensure that their needs are fully understood and met through instruction (Granger, 1994).

One of the most influential learner corpora is the International Corpus of Learner English (ICLE) compiled by Granger and her colleagues at the Université Catholique de Louvain. ICLE

is a collection of corpora produced by learners from diverse language backgrounds. The most recent version contains 3.7 million words of EFL writing from higher intermediate to advanced learners of English. The learners represent 16 different language backgrounds (Granger, Dagneaux, Meunier, & Paquot, 2009). All subcorpora of the ICLE follow the same explicit criteria throughout compiling and they are highly comparable. Many corpus studies of learner language were based on the ICLE to investigate learners' overuse and underuse of certain language aspects by comparing with native speaker English corpora (Biber et al., 2010). The ICLE is employed in the present study to examine lexical component of NNSs' academic writing and compare it to their native speaking peers'. More details of this corpus are presented in Chapter Three.

The International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2011) is another large learner corpus with learner language data from diverse mother tongue backgrounds. The corpus includes 1.8 million tokens of both spoken and written texts. The 3,550 college ELs are from 10 countries and areas in Asia. This corpus also includes a subcorpus of NS data.

There are other learner corpora focusing on learner language data from a single mother tongue background. For instance, National Institution of Information and Communications Technology (NICT) in Japanese created a Japanese Learner English (JLE) corpus which contains learner data of Japanese learners of English (Izumi, Uchimoto, & Isahara, 2004). The Ten-thousand English Compositions of Chinese Learners (the TECCL Corpus; Xue, 2015) contains Chinese English learners' written texts; and InterFra Corpus collects samples from Swedish learners of French (Bartning & Schlyter, 2004).

Compared to written corpora, oral learner corpora are much more challenging for researchers to compile. As a result, only a relatively few representative oral corpora have been compiled, such as the College English learners' Spoken English Corpus (Yang & Wei, 2005) and the Louvain International Database of Spoken English Interlanguage (Gilquin, de Dock, & Granger, 2010). The former contains only data from Chinese learners of English, while the latter contains learner data from multiple linguistic backgrounds.

Next, a few specific studies that utilized learner corpora are discussed to present the common research foci, methodology, and techniques of LCR. Examining language at a discourse level, Flowerdew (1998) compared the usages of cause and effect markers between an expert and a learner corpus. This examination revealed the overuse of logical connectors and the lack of mitigating markers (e.g., modal verbs or adverbs) in the learner corpus. This study sheds light on teaching English for science and technology purposes.

Moreover, quite a few studies analyzed subcorpora of the ICLE to investigate certain usages of learner English. For instance, based on the German component of the ICLE, Nesselhauf (2003) explored German ELs' use of verb-noun collocations. Osborne (2008) looked at the difference of adverb placement in ELs' and NSs' written texts. The language backgrounds of the learners in this study were diverse, extracted from French, Italian, and Spanish components of the ICLE. Thewissen (2013) conducted a study to see the developmental trajectories of ELs with respect to accuracy. Based on the error-tagging version of the ICLE, the researcher randomly selected learner texts from various language backgrounds, then the essays were graded and divided into various proficiency level by experts. The study intended to show the developmental patterns of L2 development based on the error analysis.

Learner corpora and studies conducted to explore teaching and learning of other L2 languages beside English are growing as well. Belz and Vyatkina (2005) presented a corpus-based intervention, Telekorp, for teaching German modal particles. The corpus used in the study was a longitudinal German-English bilingual corpus contained emails and synchronous chat between German learners of English and English learners of German over three weeks.

So far, it is not difficult to infer that comparison and contrastive analysis are the essence of LCR. Various comparisons include native and nonnative speakers' language production, longitudinal differences within the same learners across a time period, as well as similarities and differences between L1 and L2. Through comparison, researchers provide empirical and explicit support on 1) what the learners tend to lack in reaching desired proficiency; 2) the developmental progress of learners' interlanguage; and 3) potential transfer from their language backgrounds onto their L2 production. Furthermore, the empirical data from learner corpora are highly valuable for carrying out DDL activities. In next section, I further the discussion of the major methodology in the paradigm of LCR, the Contrastive Interlanguage Analysis (CIA).

Contrastive Interlanguage Analysis (CIA)

Inspired by contrastive analysis in traditional Applied Linguistics and thanks to the establishment of various learner corpora, a comparative methodological framework found its way to analyze data from learner corpora and placed emphasis on learner data naturally (Granger, 1996; Granger, 2015). First proposed by Granger (1996), the CIA has become the main method for studies conducted in the realm of LCR. The comparative method helps reveal linguistics features of the learners that may not have been easily seen if analyzed in isolation (Granger, 2015). In this section, the major components of the CIA framework are first introduced, then influential studies that applied this framework in analysis are examined next.

Then, I critically evaluate this framework by looking at its strengths, weaknesses, and existing criticisms in the field. Finally, the methodological foundation of the current study is presented by connecting the goal of the study with the appropriateness of the CIA.

Granger (1996) states that the CIA is an integrated contrastive model which combines traditional Contrastive Analysis (CA) and a new type of contrast. The distinctive feature of the CIA is that it establishes comparisons between native and learner varieties of one and the same language rather than between two different languages. As the core approach for analyzing data from the ICLE, the initial goal of the CIA is to reveal “foreign-soundingness” in learner writing (Granger, 1993). There are two types of comparison involved in the CIA framework: 1) native language (NL) vs. interlanguage (IL); 2) IL vs. IL.

With respect to the first type of comparison, by comparing to the reference corpus (i.e., NL), it distinguishes learner language by investigating errors and under- or overuse of certain language features. In terms of the second type of comparison, researchers identify the potential sources of certain non-standard features by comparing various interlanguages (Gilquin & Granger, 2015). For instance, features that are unique to one mother tongue group indicate the possible transfer from L1, while features that are common across different language groups shed light on the inherent challenges that target language learners all face. Granger (1996) developed a diagram form to represent the framework of CIA (See Figure 1).

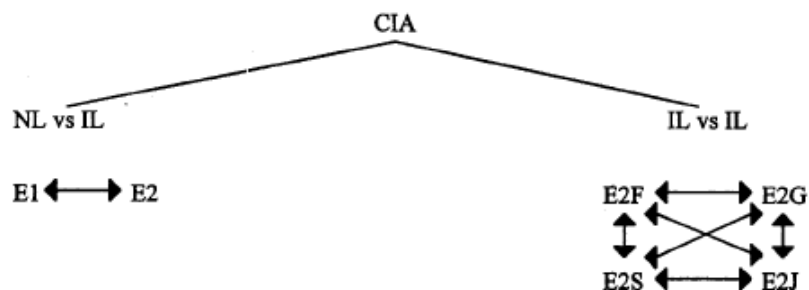


Figure 1: Contrastive Interlanguage Analysis (CIA)

Source: Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.). *Language in contrast: Papers from a symposium on text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.

In this diagram, Granger used English as an example of second or foreign language to illustrate the possible comparisons. Native-speaker English was represented as E1, learner English was represented as E2. Among interlanguages, French English learners were represented as E2F. Similarly, German, Swedish, and Japanese English learners were represented as E2G, E2S, and E2J respectively. After the CIA was developed, many studies have explicitly or implicitly based their analysis upon this framework. Here, I present a few influential studies that represent two major types of comparison in the CIA.

1) Native Language (NL) vs. Interlanguage (IL): In terms of the comparison between native language and interlanguage, Granger (1998) herself conducted a study to explore the advanced French ELs' use of prefabricated patterns, such as collocations and lexical phrases. The researcher employed the French subcorpus of the ICLE as the interlanguage corpus and three native speaker corpora as the reference corpora. The three NS corpora included the Louvain essay corpus, the student essay component of the International Corpus of English (ICE), and the Belles Lettres category of the Lancaster-Oslo/Bergen Corpus (LOB). In terms of collocations, the comparison results showed the learners' underuse of native-like collocations and common use of atypical word combinations. Moreover, the learners' responses to word-

combination test indicated their weaker sense of collocations than the NSs. For lexical phrases, the analyses showed the French ELs' similar use of passive structured prefabs and overuse of active structured prefabs when compared to those within native texts. Thus, the researcher pointed out that the possible reason of "foreign-soundness" can be ascribed to both under- and overuse of prefabs.

There are many other studies that also compared various aspects of differences in language usage between NS and NNS corpora. For instance, Altenberg and Tapper (1998) and Eia (2006) respectively found that Swedish and Norwegian learners tended to excessively use informal connectors, such as *and* and *but*, while the use of formal connectors was relatively underused when compared to NS data. Furthermore, Ishikawa (2013) conducted a contrastive analysis based on the International Corpus Network of Asian Learners of English (ICNALE) and the parallel NS corpus included. The analysis involved Japanese EFL learners' most frequently overused words compared to NS corpus and the contribution from proficiency levels. The results suggested that Japanese learners tended to overuse indefinite personal nouns or pronouns, such as *we* and *people*. Meanwhile, thought-related verbs (e.g., *agree* and *think*) were overused by Japanese learners as well. Compared to NSs, Japanese ELs also used the modal verbs of obligation more often. All in all, the application of the CIA greatly facilitates the identification of certain language features of learners' interlanguages.

2) Interlanguage (IL) vs. Interlanguage (IL): With regard to the comparison between interlanguages, Osborne (2008) looked at the differences between Romance L1 ELs and Germanic L1 ELs in the placement of adverbs. Of particular interest to the study was the placement of adverbs in verb-adverb-object order, which is not considered as a norm in modern English but often occur in learner English on the post-intermediate level. The results showed a

strong tendency of using verb-adverb-object order among Spanish, Italian, and French learners of English. However, other language speakers were observed with less frequent use of this specific semantic order; the fact that those speakers' L1s are more reluctant to place adverbs in this position might be the explanation for this phenomenon.

Paquot (2010) proposed the Academic Keyword List (AKL) and employed the CIA to test the validity and reliability of the list. Paquot used subcorpora from the ICLE with 10 different mother tongues and the academic component of the BNC to examine the differences of NSs' and NNSs' usage of academic vocabulary. The wide range of learner language backgrounds allowed the interlanguage features to be represented clearly; meanwhile, different interlanguage features based on the first languages were revealed.

In sum, the second type of comparison in the CIA helps researchers focus more on the varieties of learner output and explore the potential contribution from L1, thus emphasizing the deeper features of interlanguages.

Although the CIA has been widely accepted in the field of LCR, criticism has also risen towards the beliefs behind the framework. Two major criticisms include the notion of "comparative fallacy" and the concept of the norm (Granger, 2015). Both issues are more or less related to the sociolinguistic perspective in SLA. The notion of "comparative fallacy" states that comparing learner interlanguages to the so-called target language contributes to the constant deficient position of interlanguages, which may seriously hinder learners' language development (Bley-Vroman, 1983). The second criticism is related to the increasing attention toward World Englishes and English as a Lingua Franca. When discussing the usage of native language, the concept of norm is becoming controversial (Brutt-Griffler & Samimy, 2001). Tan (2005) used

the emerging variety of Thai English as an example and emphasized that without considering the local environment, imperialistic assumptions about the ownership of English may result.

In response to the criticisms, Granger (2009, 2015) acknowledges the underlying idea behind it. In fact, she even regards the criticisms as an incentive for moving the field forward. However, in her rebuttal, she justified the framework by explaining its original purpose and pedagogical function. First, Granger (2009) argues that most proficiency assessments in the field of SLA are established on the underlying L1 norm in order to evaluate learners' answers and performances. Lardiere (2003) adds that the norm can be considered as a legitimate starting point for learners and teachers to be aware of the deviation from the target language. The issue that needs attention is what kind of conclusion should be made when the learners' language product varies from the target language norm. Furthermore, from the pedagogical viewpoint, knowing what the learners do right or wrong, or even partly wrong gives the teachers sufficient support while designing activities or lesson foci (Granger, 2015).

Second, in terms of the controversial concept of the native norm, Granger (2015) argues that the CIA does not hinder the diverse varieties of reference corpora. As a matter of fact, Granger (1998) provides a wide range of choices of native corpora from the International Corpus of English (ICE). Currently, there are 13 varieties of native corpora available from the ICE. Also, the description of learner language usage in terms of under- or overuse of certain language features is neutral and merely descriptive not prescriptive (Gilquin & Paquot, 2008). On the other hand, Granger (2015) admits that it is necessary to reconsider the dichotomy of native and nonnative concepts. She states that the term native and nonnative should be avoided as *de facto* generic terms in the CIA. In the current study, the potential harms of considering NSs' writing products as the norm, even the faultless variety is well acknowledged. The conduction of the

present study intends to demonstrate descriptive profiles of learner language rather than claiming that NSs' writing is better than that of NNSs. Keeping this notion into consideration, words like native and nonnative writing are still used throughout this dissertation to avoid misunderstanding and complication.

In reappraising the CIA, Granger (2012) included the notion of expert variety in addition to native variety when redefining the framework. Moreover, Granger (2015) designed a new version of the CIA, CIA², which is more inclusive and comprehensive (See Figure 2).

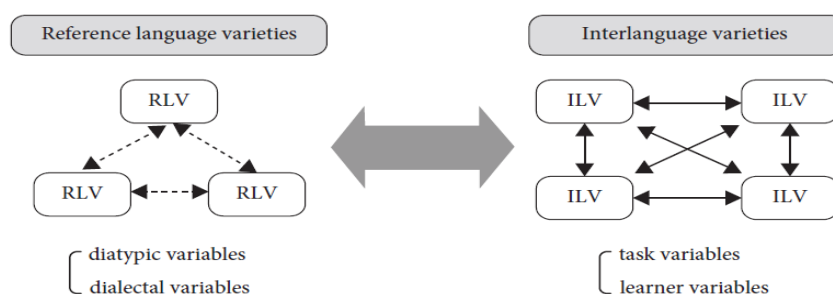


Figure 2: CIA²

Source: Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24.

In the new version, the notion of “varieties” is emphasized and refers to both reference language varieties and interlanguage varieties. The multiple corpora included in the category of Reference Language Varieties (RLV) indicate the possibility of employing both inner and outer circle varieties of the target language (Granger, 2015). Meanwhile, possible comparisons can be conducted between the reference varieties in terms of diatypic and dialectal variables. The employment of the word “reference” indicates that the corpora included are not necessarily a so-called norm. The category of Interlanguage Varieties (ILV) introduces the potential comparison between both task variables and learner variables. In particular, the recorded metadata in many learner corpus projects are neglected in the analyses. Therefore, the CIA² is not a brand-new framework, rather it is highly built upon the original structure of the CIA. By updating the

framework and including more varieties, it provides a promising path to achieve deeper understanding of learner language.

One main aim of the current study is to examine the possible differences between language learners and their native-speaking peers in terms of lexical quality in academic writing. The other research question deals with the differences between various learner corpora to see the potential influence from L1 or educational background. Therefore, both examinations will be mainly carried out through comparisons between 1) learner corpora and NS corpora; 2) all subcorpora of English writing of writers from various mother tongue backgrounds. CIA² is adopted as the methodological framework in the present study because of its broader notion of language products and the multiple NS and learner corpora.

Summary

The review of the scholarship in this section establishes the theoretical and methodological foundation of the present study. Regarding the theoretical foundation, the comprehensive and authentic features of corpus linguistics provide a solid base for extracting patterns of language usages from multiple corpora; meanwhile, the advancement of technology relatively eases the examination of a large number of texts. Therefore, it is appropriate to explore the current research questions from a corpus-based perspective.

In terms of methodological foundation, the present study aims to uncover the lexical features of interlanguage and compare them to NSs' performance. The second version of CIA, CIA², is employed as the methodological foundation to guide the investigation.

Next, the specific lexical aspects in academic writing that the present study examines are discussed. Corresponding approaches of measurement are introduced.

Lexical Features and Corresponding Measurements

The previous two sections of this chapter introduced the rationale that the present research questions built upon. In a nutshell, writing is an essential skill that almost every L2 learners in an academic setting needs to develop and vocabulary plays a critical role in developing L2 writing skills. Thus, analyzing vocabulary usage in academic writing in the present study is of great value. In terms of how to analyze the vocabulary aspect in academic writing, corpus-based CIA² framework is used in the current study owing to its validity, reliability, and efficiency. In this section, I further the discussion to specify the target lexical features in academic writing that are examined in the present study and the corresponding measuring approaches.

As described in the former sub-section of “Vocabulary and Writing,” research has achieved a relatively solid conclusion that vocabulary competence significantly correlates with holistic writing quality. Thoroughly searching the body of literature, there are a large number of studies assessed the participants’ receptive or productive vocabulary knowledge through vocabulary tests to evaluate their vocabulary level. Most of these studies analyzed the relationship between the participants’ vocabulary levels and their writing products. However, one obvious weakness of this type of research is the limited number of participants. For instance, Stæhr’s (2008) study recruited a comparatively large number of participants ($N = 88$), which still has less than 100 participants and the total number of tokens is less than 50,000. Hence, in spite of the validity and reliability of the studies, it is not entirely convincing to generalize their findings to different population groups given the diversity of L2 learners.

Regarding this issue, some studies included multiple corpora with learner and/or NS data to enlarge the generalizability of the findings. For instance, Paquot (2010) used subcorpora from

the ICLE with 10 different mother tongues and the academic component of the BNC to examine the differences of NSs' and NNSs' usages of academic vocabulary. Nevertheless, lexical component in writing is a multi-dimensional construct rather than merely frequency. Lexical diversity, lexical errors, and collocations are a few examples that all contribute to lexical quality in writing. To date, very few studies have compared lexical features with a large number of NS and NNS writing samples in a comprehensive manner. Thus, the present study aims to address the two aforementioned gaps in the literature: 1) using corpus data to present the generalizable patterns of lexical usage in both NESs' and NNEs' writing; 2) including various lexical features to present a global picture of the position of vocabulary in writing and the differences between the performances of NESs and NNEs in the lexical component of academic writing.

In this section, I first provide definitions of commonly used terminologies in the field of Corpus Linguistics and various notions of lexical features. Later, I present detailed explanations of the selected lexical features and the corresponding measurements.

Major Concepts and Definitions in Corpus Linguistics

Before furthering the discussion to specific measures of lexis, some commonly used terminologies in Corpus Linguistics are worth noting here. *Tokens* and *running words* both refer to the number of words in a text; *types* refers to the number of different words in a text. A *type* is also called an individual *word form*. For instance, the sentence, "the white cat is bigger than the black cat" contains nine tokens, namely nine running words; however, there are seven types (word forms) in the sentence, namely, *the*, *white*, *cat*, *is*, *bigger*, *than*, *black*. A *lemma* is a group of word forms that are related by being inflectional forms of the same base word. For instance, the verb *destroy* can be considered as a base word; its inflected forms, including *destroys*, *destroying*, and *destroyed*, are all part of the verb lemma. They are considered as one lemma.

However, the noun form *destruction* is a separate lemma due to the derivational rather than inflectional form changing. Different from the concept of *lemma*, the concept of *word family* includes both inflectional and derivational word forms (McEnery & Hardie, 2012). For instance, in the above example, all five word forms, *destroy*, *destroys*, *destroying*, *destroyed*, and *destruction* are considered as one word family.

Crossley and McNamara (2009) note that to truly understand L2 learners' lexical proficiency, measurements on both surface and cognitive levels are needed to provide a comprehensive and profound description. Surface level measuring of lexical features mainly deals with the frequency and characteristics of certain groups of words, which can be explained by a plain interpretation. Some common terms used in the surface level measuring include lexical richness, sophistication, and density. Various scholars have employed different classification methods and definitions of these terms according to the needs of their studies. Thus, there has yet been a complete agreement on the definitions of the above terms. In the next section, I introduce the popular perspectives regarding how to define various terms related to surface level measuring of lexical features; meanwhile, I justify my perspective and choice by considering the characteristics of the present study.

With respect to cognitive level measuring of lexical features, it mainly addresses the learners' deeper sense of words in association with the syntagmatic and paradigmatic properties (Crossley & McNamara, 2009). In other words, the cognitive level of lexical usage discusses how a single lexical unit relates to a larger context to reach the purpose of the writing. Therefore, cohesion level in writing is the major component of measuring lexical features at the cognitive level. The construct of cohesion is explained shortly in the following sections.

Major Lexical Features in Writing

In previous research, various lexical features have been studied. This section introduces major lexical features appeared in prior research. These lexical features are categorized into surface and cognitive levels.

Surface Level Lexical Features

In a broad sense, lexical proficiency in writing is widely considered as a writer's ability in using different levels of vocabulary appropriately (Crossley, Salsbury, McNamara, & Jarvis, 2011). It is commonly assumed that the more high-level vocabulary is used in a text, the more advanced a writer is (Laufer & Nation, 1995). Here, surface level measuring of lexical proficiency is closely related to the frequency of the vocabulary. Despite various definitions and classification methods, the coverage of certain types or levels of vocabulary in a text is the key approach for measuring.

Polio (2001) states that surface level measures of lexical proficiency include lexical originality/individuality, sophistication, diversity, density, errors, and diversity of form class. Lexical originality or individuality indicates the relationship between an individual writer and a group. Laufer (1991) notes that lexical originality is defined as the percentage of lexemes that are included in an individual's writing but not in other group members' writing. Therefore, the value of lexical originality is unstable and unreliable, it changes when a text is compared to different groups of texts (Laufer & Nation, 1995).

Lexical sophistication refers to the coverage of advanced vocabulary in a text (Engber, 1995; Laufer & Nation, 1995). In Laufer and Nation (1995), an approach to evaluate lexical sophistication level was developed as the Lexical Frequency Profiles (LFP), which is explained in detail shortly.

The major principle to assess lexical diversity is to evaluate the different types of vocabulary compared to the tokens in a text. Laufer and Nation (1995) calculated lexical diversity through the process of dividing the number of types by the quantity of tokens. However, there has been criticisms toward the simple ratio between types and tokens due to the influence of text length. Thus, other measures have been developed to eliminate the effect of text length. Detailed explanations toward various measures of lexical diversity are provided shortly in the following review.

Lexical density refers to the percentage of lexical words compared to the total number of words, namely the combination of lexical and functional words (Laufer & Nation, 1995). Laufer and Nation also note that the value of lexical density does not necessarily reveal the quality of lexis used in the text since it varies much with the change of syntactic and cohesive properties.

Lexical error is another feature revealing the surface level measures of lexical proficiency. The ratio between the number of lexical errors and the total number of errors is often used to represent the level of lexical errors. The analysis of lexical errors requires the use of accurate and error-annotated corpora as well as qualitative approaches, both of which require a vast amount of time and resources.

Finally, the diversity of form class refers to the ratio between nouns and the total number of lexical words; verbs and the total number of lexical words; adjectives and the total number of lexical words; and so forth. Similar to lexical density, the use of various parts of speech differs greatly with the change of topics and writing styles. Table 5 summarizes the definitions of the major terms explained by Polio (2001).

Table 5

Surface Level Measures of Lexical Features (Polio, 2001)

Term	Explanation	Stability
Lexical originality (individuality)	The relationship between an individual writer and a group	No. Changing when a text is compared to different groups
Lexical sophistication	The coverage of advanced vocabulary in a text	Yes.
Lexical diversity	Evaluating the different types of vocabulary compared to the total number of tokens in a text	Depending on different measurements.
Lexical density	The percentage of lexical words relative to the total number of words, namely the combination of lexical and functional words	No. Varying with the change of syntactic and cohesive properties
Lexical errors	The ratio between the number of lexical errors and the total errors.	Requiring standardized error classification, error-annotated corpus.
Diversity of form class	The ratio between nouns and total number of lexical words, verbs and total number of lexical words, adjectives and total number of lexical words, etc.	No. The use of various parts of speech differs greatly with the change of topics and writing styles.

Another widely accepted classification of lexical features derives from Read (2000), where lexical richness is considered as the major component for evaluating lexical proficiency in writing. In Read's (2000) notion, lexical richness is a cover term including four facets: lexical variation (diversity), lexical sophistication, lexical density, and lexical errors. In terms of lexical variation (diversity), density, and sophistication, the definitions are similar as in Polio (2001).

Figure 3 depicts Read's conceptualization of lexical richness.

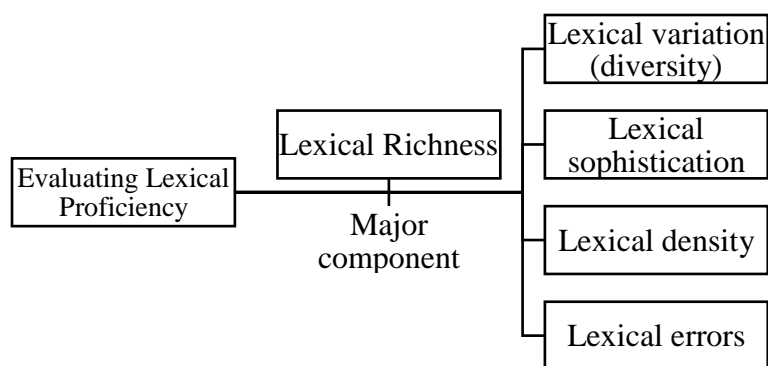


Figure 3: Lexical Richness

Regarding lexical errors, Read (2000) emphasizes that the concept is not only related to accuracy but also appropriateness. In addition, the classification scheme that scholars develop for lexical errors can be highly subjective. To sum up, lexical richness is an inclusive concept which consists of not only vocabulary size but vocabulary depth of knowledge, meaning how well a writer can apply a particular word in discourses (Nation, 2001).

Many studies that investigated lexical quality of texts based their measures on Read's (2000) notion of lexical richness, which provides a comprehensive perception of how lexis influences the holistic quality of a text. However, Jarvis (2013) recently proposed a change toward lexical diversity in the construct of lexical richness. Jarvis did not question the theoretical construct of lexical diversity; nevertheless, he considered the issue from a mathematic and statistic viewpoint. He argues that the measuring of lexical diversity is not merely frequency related; in order to achieve the most accurate measurement of lexical diversity, researchers should go beyond the simple equation of dividing types by tokens. Therefore, since the measurement of lexical diversity is so different from measuring other indices in lexical richness, Jarvis proposes that lexical diversity should not be included in lexical richness. Rather, lexical diversity should be considered as an independent component that influences the holistic lexical quality of a text.

Synthesizing the existing perceptions of measuring lexical features, the concept of lexical richness from Read (2000) is more comprehensive and well-organized, which provides an inclusive understanding of the key indices. A few indices stated in Polio (2001) are not included in lexical richness by Read (2000), such as lexical originality and diversity of form class. Both of these items vary largely when the target corpus or topics change. Thus, the present study eliminates the measuring of lexical originality and diversity of form class. The measure of lexical density also varies depending on different texts. Hence, lexical density is eliminated in the present study as well. Finally, due to the large amount of time and resources required to annotate the corpus data as well as subjective classification of errors, lexical errors is beyond the scope of the present study.

To sum up, adopting and modifying Read's (2000) notion of lexical richness, the present study examines lexical diversity and lexical sophistication as the surface level measures in both native and nonnative speakers' academic writing.

Cognitive Level Lexical Features

Different from surface level measures, cognitive level measures of lexical quality have not earned sufficient attention and empirical analysis. The concept of cognitive level measures of lexical features is mainly studied by Scott Crossley, Danielle McNamara and their colleagues. Crossley and McNamara (2009) explicitly emphasized the deeper insights that cognitive measures of lexis can provide for understanding how learners process and produce a second language. As mentioned before, vocabulary depth of knowledge contributes significantly to the holistic writing quality; in this case, the measure of one's cognitive sense of lexis and the connection between single vocabulary units largely indicate the learner's vocabulary depth of knowledge.

Since not too many different strands exist in cognitive level measures, I build my analysis upon Crossley and McNamara's (2009) study to examine the cohesion levels of native and nonnative English writers' academic writing. Later, I further explain the value of lexical quality at the cohesion level in academic writing and its corresponding measurement.

Target Lexical Features and Measurements

In this section, three measures of lexical features described above are further explained and their corresponding measurements are introduced. Since writing is a cognitive process regardless what frequency level the vocabulary is, to avoid misinterpretation, the present study eliminates the division of lexical features by using the dichotomy of surface and cognitive levels. The three target lexical features that were measured in the present study include lexical diversity, lexical sophistication, and cohesion.

Lexical Diversity

Definition

Lexical diversity has been defined and researched from various perspectives in the literature. Some have used it interchangeably with vocabulary richness (Wimmer & Altmann, 1999), lexical variation (Granger & Wynne, 2000), and lexical density (O'Loughlin, 1995). Among various notions of lexical diversity, some focus on the number of different words used in a text while others emphasize on the difficulty or relative rarity of the words used. Laufer (2003), from the perspective of L1 development, defines the concept of lexical diversity by combining the percentage of infrequent vocabulary and the percentage of different words in a composition. However, Read (2000) took the perspective of L2 learning and synthesized the multidimensional concept of vocabulary richness, including lexical variation, lexical sophistication, lexical density, and the number of errors. In Read's definition, lexical variation equals lexical diversity, which

refers to the range of vocabulary and avoidance of repetition. It is measured by comparing the number of different words with the total number of words written, which is traditionally the type-token ratio. Because Read's (2000) perspective focuses on the variation level of the vocabulary used in the writing and is more related to L2 writing, his definition of lexical diversity is adopted in the present study.

The relationship between lexical diversity and holistic quality of academic writing

Various research has examined and compared the extent of lexical diversity in academic writing of NSs and NNSs as well as its contribution to the holistic quality of the writing. Under different research conditions and measurements, most empirical studies have come to a congruent perception: higher lexical diversity correlates to high quality of the writing, while lower variation and diversity of vocabulary in a text indicates lower quality of the text in a holistic way (Eckstein & Ferris, 2018; Ferris, 1994; Friginal, Jarvis, 2002; Li, & Weigle, 2014; Silva, 1993; Yu, 2009). Table 6 summarizes the empirical studies that reveal the relationship between lexical diversity and the holistic writing quality.

Table 6

Lexical Diversity and Holistic Writing Quality

Study	Main Findings
Eckstein & Ferris, 2018	- Significant difference in lexical variation between L1 and L2 students' writing.
Ferris, 1994	- L2 writers had a smaller lexical repertoire than their L1 peers.
	- More advanced learners were able to employ a wider variety of lexical choices, syntactic constructions, and cohesive devices than did those at a lower level proficiency.
Friginal et al., 2014	- Highly rated essays made use of a wider range of vocabulary than did the lower level essays.
	- NS essays had a higher possibility (43%) of containing long text length and high lexical diversity than did NNS essays (33%).
	- The "expert" NNS learners were able to produce longer and more lexically diverse texts and received higher rating.
Jarvis, 2002	- Significant difference of lexical diversity between NNS and NS corpora ($p < .01$).
	- NSs tended to produce higher levels of lexical diversity than NNSs did.
Yu, 2009	- Lexical diversity has substantial and significant correlations with the producers' writing abilities and overall language proficiency.
	- Other factors such as gender, L1 background, test taking purpose, and writing prompt also contribute to the holistic quality of the composition.

Measurements

A number of measuring methods and techniques of lexical diversity have been developed and utilized in previous studies. Results of these studies indicate how lexical diversity is conceptualized and how it is measured largely influence the final result (Jarvis, 2002). Three major methods have been employed to measure lexical diversity: type-token ratio (TTR), the *D*, and the Measure of Textual and Lexical Diversity (MTLD).

First, TTR is the traditional method to quantify lexical diversity (Laufer & Nation, 1995; McCarthy & Jarvis, 2007). Here, lexical diversity is represented by the ratio between types and tokens. De Haan and van Esch (2005) compared Spanish and English learners' essays to those of their native speaking peers to explore the lexical features of the writing samples in two

consecutive years. TTR was used to measure the level of lexical diversity of the essays.

Although the findings confirmed the increasing TTR of Spanish learners' essays throughout two years and the higher TTR of native writers' essays, the results showed a reverse trend in nonnative English writing. Thus, the authors suggest cautious use of TTR in analyzing lexical diversity as it may lead to ambiguous results. Additionally, TTR is highly influenced by the text length, which may contribute to unreliable results if used without consideration (Laufer & Nation, 1995; McCarthy & Jarvis, 2010).

Second, Durán, Malvern, Richards, and Chipere (2004) devised a program, *vocd*, to calculate lexical diversity by using the mathematical model of the *D*. The *D* is believed to be independent from the limitation of text length. The principle behind this measurement is that the system randomly samples 35 to 50 tokens from a text 100 times to form a theoretical curve. A TTR score is calculated for each of the samples to produce a mean score that acts as the *D*-score for each sample. Then, all of the *D*-scores are averaged to reach a further mean. In the end, the above procedure repeats three times and a final mean of *D*-score is produced as the rating of lexical diversity. The *D* was tested in the same study with 32 children's speech samples over the study. The results of the study suggested increasing *D* scores with advancing age of the participants. Jarvis (2002) also suggests the accurate curve and consistency that the *D* formula provides, stating that the *D* index is optimal for comparing texts of different lengths. However, McCarthy and Jarvis (2007) used *vocd* to further test the validity of the *D* value for representing lexical diversity. The results indicate that the *D* measurement is only reliable for texts with low lexical diversity, for instance, children's or NNS learners' discourses.

Finally, the most recent method of measuring lexical diversity is using the MTL_D value, which is believed to be able to solve the accuracy and reliability limitations that the *D*

measurement contains (McCarthy & Jarvis, 2010). The operating principle of the MTLD is that the TTR scores continuously decline from 1 to 0 as the text progresses. The calculation of the MTLD index makes use of a notion closely related to thematic saturation to achieve the precise point of lexical diversity decline. After finding the reliable TTR point (.72), the MTLD counts the number of times the TTR value occurs in the text, then divides the result by the total number of tokens. McCarthy and Jarvis state that the average score is between 70 to 120 with 120 as the highest level of lexical diversity. The last step is to run the entire process backward to check the accuracy of the cutoff point. Table 7 summarizes the major characteristics of the above measurements for lexical diversity.

Table 7

Measurements of Lexical Diversity

Measurements	Key Features
Type-token ratio (TTR)	<ul style="list-style-type: none"> - Traditional method - Highly influenced by the text length, may contribute to unreliable results if used without consideration.
Vocd	<ul style="list-style-type: none"> - Using the mathematical model of the <i>D</i>. - The <i>D</i> is believed to be independent from the limitation of text length. - Increasing values of <i>D</i> with advancing age of the participants. - The <i>D</i> measurement is only reliable for texts with low lexical diversity, for instance, children's or NNS learners' discourses.
MTLD	<ul style="list-style-type: none"> - Believed to be able to solve the accuracy and reliability limitations that the <i>D</i> measurement contains. - A robust measurement for lexical diversity without being influenced by text length - More studies need to be conducted to validate the effectiveness of the MTLD in analyzing lexical diversity of texts from different registers and with various text lengths.

Admittedly, more studies need to be conducted to validate the effectiveness of the MTLD in analyzing lexical diversity of texts from different registers and with various text lengths.

Compared to TTR and the *D*, theoretical explanations and preliminary studies have supported the

reliability of the MTLD as a robust measurement for lexical diversity without being influenced by text length (McCarthy & Jarvis, 2010). Hence, the present study employs the MTLD to measure the lexical diversity of the texts. The Coh-Metrix 3.0 (<http://cohmetrix.com/>) is a free online program that calculates the MTLD value of a text.

Lexical Sophistication

Definition

Laufer (1994) and Laufer and Nation (1995) consider lexical sophistication to be one of the measures for lexical richness. Both pieces of scholarship define lexical sophistication as the percentage of advanced words in texts. Read (2000) follows this perception and defines lexical sophistication as “the use of technical terms and jargon as well as the kind of uncommon words that allow writers to express their meanings in a precise and sophisticated manner” (p. 200). Thus, concluding the mainstream conceptualization of lexical sophistication, frequency is the principle factor that determines whether the lexical items are sophisticated or not. The common mechanism is that the more sophisticated a lexical item is, the less frequent it occurs in use.

The relationship between lexical sophistication and holistic quality of academic writing

The high level of lexical diversity contributes to the more sophisticated use of vocabulary. As another key component of lexical richness, lexical sophistication suggests the writer’s ability to employ less frequent vocabulary. Research studies support the positive correlation between level of proficiency and lexical sophistication (Laufer, 1994; Laufer & Nation, 1995; Silva, 1993). In addition, Muncie (2002) analyzed the English vocabulary development of process writing of 30 Japanese university students. The results from the final drafts suggested a significant higher percentage of sophisticated words. Similar findings indicate the positive effect of using sophisticated words for improving academic writing quality (Kormos,

2011; Kyle & Crossley, 2015). Table 8 summarizes the major findings regarding the correlation between lexical sophistication and holistic writing quality.

Table 8

Lexical Sophistication and Holistic Writing Quality

Study	Main Findings
Kormos, 2011	- L2 writers ($n = 44$) used more high-frequency words than native English writers ($n = 10$).
Kyle & Crossley, 2015	- Words used in fewer contexts are considered to be more sophisticated than those that are commonly used.
Laufer, 1994	- Lexical sophistication is highly correlated with holistic writing quality
	- As ELs' proficiency level increased ($N = 48$), their use of high-frequency words decreased, the use of low-frequency words increased, the use of academic and off-list words increased.
Laufer & Nation, 1995	- ELs at different proficiency levels presented significantly different lexical profiles: advanced learners used more academic and off-list words and fewer high-frequency words; learners with lower proficiency level demonstrated the opposite profiles.
Muncie, 2002	- Compared to first drafts, same EL population ($N = 30$) presented increased lexical sophistication level in their final drafts and higher holistic writing quality.

Measurements

After raising the definition of lexical sophistication, Laufer (1994) and Laufer and Nation (1995) challenged the uncertainty in classifying “advanced” words. Laufer (1994) states that whether a lexical item can be considered as advanced largely depends on the learners' proficiency level. For instance, a lexical item may be considered as sophisticated for a beginning level learner but not for an advanced level learner. Thus, to avoid unilateral judgement, Laufer (1994) and Laufer and Nation (1995) proposed the construct of Lexical Frequency Profiles (LFP) as a comprehensive method to measure the lexical sophistication of a text. The LFP determines whether the lexical usage of a text is sophisticated or not based on four frequency bands. The first and second frequency bands are the first 1,000 and second 1,000 most frequent words in

English, and the ratings show the coverages of the first 1,000 and second 1,000 words in the text respectively. The third frequency band is from Xue and Nation's (1984) University Word List (UWL), which consists of 836 academic words. The LFP generates the third coverage, namely the proportion of these 836 academic words in a text. The new version of the LFP employs the Academic Word List (AWL; Coxhead, 2000) for the third frequency band. Finally, the last frequency band is the words that are excluded from the first three bands, and a corresponding coverage is also calculated (Laufer, 1994; Laufer & Nation, 1995). The program of LFP is freely accessible on the Compleat Lexical Tutor website (Cobb, n.d.).

Since the establishment of the LFP, multiple studies have employed this measurement and tested its validity and reliability. Laufer (1994) tested the validity of the LFP in her own study of 48 undergraduate Israeli EFL learners whose lexical growth was examined by the LFP. The results confirmed that learners were able to use more low-frequency words as their proficiency progressed. Later, Laufer and Nation (1995) conducted another study to confirm the validity of the LFP. In this study, writing products by three proficiency levels of English learners were examined by using the LFP. The results indeed showed that the learners demonstrated stair-step lexical profiles. Among the three proficiency groups, majority of the words in the texts were from the first 1,000 frequency band. Additionally, learners with higher proficiency were able to use more low-frequency words. Subsequent studies conducted by Nation and Waring (1997) and Valcourt and Wells (1999) also support the reliability of the LFP.

However, some researchers questioned the validity, sensitivity, and employment of raw frequency in accurately demonstrating the lexical sophistication of a text. For instance, Meara (2005) argued that the LFP was not sensitive enough to demonstrate modest changes in vocabulary size. In his argument, Meara used a set of Monte Carlo simulations generated by

computer to evaluate the main claims made by Laufer and Nation (1995) about the LFP. The results did not support the claims made by Laufer and Nation regarding the robustness of the LFP in distinguishing learner groups at different levels of proficiency. Laufer (2005) later also made a convincing rebuttal against Meara's criticism and questioned the value of Monte Carlo simulations for real world language usage; additionally, Laufer claimed that the purpose of the LFP is to present the productive vocabulary use rather than vocabulary size as Meara (2005) interpreted. Each side holds their own argument and the result is inconclusive.

Schmitt (2010) discussed the reliability of the LFP as well. First, after the AWL was included as the academic words band, he raised the issue that the AWL is not entirely frequency based with some words being extremely frequent but others not. Thus, it is not appropriate to consider the profiles generated by the LFP to be in a sequential order. In this case, Schmitt proposed to interpret the LFP from three levels: first and second 1,000, and others (AWL and Not on Lists). However, this would make the division rather crude and difficult to present the modest differences between learners' output. Second, Schmitt (2010) was concerned about the degree of mastery that the LFP can indicate. Indeed, the LFP may present the coverage that certain frequency band of vocabulary takes in a text; however, it has no information about whether the vocabulary is used appropriately. This was similar as one of the questions that Meara (2005) raised, which is how the LFP copes with incorrect or inappropriate use of lexis.

Most recently, Kyle and Crossley (2016) extended the indices to measure lexical sophistication beyond frequency. They argue that word range, bigram and trigrams (i.e., two- and three-word strings), academic words and phrases, psycholinguistic properties of words, the semantic relationships all contribute to the understanding of lexical sophistication and L2 writing performance. Their study investigated the relationship between these newly developed indices of

lexical sophistication and holistic writing quality in both independent and source-based writing tasks. The results support the strong influence of word range and bigrams on the writing quality of independent tasks. The automatic analysis tool that Kyle and Crossley developed is the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015). Since there has been not sufficient studies to evaluate the validity of the tool and it is not yet widely accepted in the field, thus the present study does not evaluate and employ this new tool to examine lexical sophistication levels of the target corpus data. However, once its validity is verified, this tool could certainly be considered for future research.

Overall, it can be seen that even with a few limitations, the LFP is still considered as the main tool for examining lexical sophistication of texts. As a result, this study adopts the LFP to demonstrate lexical sophistication levels of the target texts from the corpus data.

One adjustment is made to improve the validity of the LFP in examining the component of academic words. As Schmitt (2010) questioned the appropriateness of the third band of academic words by using the AWL, other scholarship has also challenged the value of the AWL in the context of discipline-based language teaching (Hyland & Tse, 2007), vocabulary teaching and researching (Nation & Webb, 2010; Paquot, 2010). However, since the use of academic words is an important component of demonstrating lexical sophistication, therefore, another list of academic words, the Academic Vocabulary List (AVL), is selected to complement the AWL.

As of now, Gardner and Davies' Academic Vocabulary List (2013) is the most recent academic word list that is compiled systematically and has been validated through large corpus data. The AVL was compiled from the academic subcorpus of the 425-million-word Corpus of Contemporary American English (COCA) (Davies, 2008-). The academic subcorpus of COCA contains 120 million running words in texts published in the United States and nine disciplines

are covered. The words that can be included in the AVL have to meet the following criteria: 1) at least 50% more frequent in the academic subcorpus than in the non-academic portion of COCA to eliminate general high-frequency words; 2) occur with at least 20% of the expected frequency in at least seven of the nine academic disciplines; 3) do not occur more than three times as often as the expected frequency in any specific disciplines. The last two criteria are set to exclude technical words and words that are only frequent in one or two disciplines. In sum, highly frequent non-academic words and highly frequent discipline-specific words are both excluded from the AVL (Gardner & Davies, 2013).

Another distinguishing feature of the AVL is that it lists lemmas rather than word families to present a more pedagogy-oriented word list. Lemmas are words with a common stem, same part of speech, and related by inflection only. For instance, the verb *assign* can be considered as a base word; its inflected forms, including *assigns*, *assigning*, and *assigned*, are all part of the verb lemma. They are considered as one lemma. However, the noun form *assignment* is considered as another lemma due to the different part of speech from the verb *assign*. The verb *assign* and the noun *assignment* are included in one word family. Using lemmas to compile word lists is preferable from the pedagogical point of view (Schmitt & Zimmerman, 2002). In their study, Gardner and Davies (2013) also tested the coverage of the AVL. The result showed that it covered about 14% of academic sections in both COCA and the British National Corpus (BNC), which indicated a higher coverage than the AWL (7.2% in COCA and 6.9% in BNC).

Olsson (2015) used both the AWL and the AVL to investigate Swedish English learners' academic vocabulary usage in writing. The result indicated that compared to the AWL, the high coverage of the AVL was able to provide more detailed description of students' writing progress.

The AVL is currently available as an embedded part of an online resource found at www.wordandphrase.info/academic. The site allows users to enter textual data and obtain frequency information of the AVL words. Therefore, to have a precise comparison and achieve relevant pedagogical implications, the AVL is used as another reference word list to examine the coverage of academic vocabulary in the texts to complement the AWL coverage.

In sum, the LFP has been selected in the current study to measure the lexical sophistication of the texts, including the coverage of the first 1,000, second 1,000, the AWL, and off-lists words. In addition, to present a more comprehensive and reliable academic lexical usage, coverage of the AVL is also included to complement the description of lexical sophistication.

Cohesion

Definition

Textual cohesion plays a critical role in connecting ideas in a text and helping readers comprehend the content with less disruption. Louwerse (2004) distinguishes the concept of cohesion from coherence. He notes that coherence is related to the consistency of the text from the perspective of readers' mental process; while cohesion refers to the elements of the text that indicate the coherent feature of the text. Cohesive devices are common elements that contribute to the cohesion of a text. Therefore, the measure of cohesive devices is a major indication of the cohesion level of a text (Crossley & McNamara, 2009; Hinkel, 2001). Normally speaking, more cohesive devices involved in a text indicate high coherent level and easy comprehension characteristics of the text.

Halliday and Hasan (1976) suggest five major categories of cohesion, including substitution, ellipsis, reference cohesion, conjunctive cohesion, and lexical cohesion. Among

these five classes, substitution and ellipsis are often used in spoken discourse; other three are more common in written discourses. Reference cohesion often present as pronominals, demonstratives, and definite articles. Conjunctive cohesion often refers to the conjunctives in writing. The major role of conjunctives can be to describe an additive, adversative, causal, temporal, or continuative relationship. Finally, lexical cohesion refers to the connective meaning that a lexical item possesses rather than the outside relationship it makes, repetitions and synonyms are examples of lexical cohesion. Many studies have adopted Halliday and Hasan's framework to examine the relationship between textual cohesion and holistic quality of the writing.

The present study focuses on the evaluation of cohesion in native and nonnative English academic writing from a quantitative perspective. Graesser, Crossley, McNamara and their colleagues developed the online tool, Coh-Metrix, to measure cohesion from a deeper level beyond cohesive devices (Graesser, McNamara, Louwerse, & Cai, 2004). This measurement tool is introduced shortly.

The relationship between cohesion and holistic quality of academic writing

Ferris (1994) analyzed 28 lexical and syntactic features of 160 ESL texts of four language groups. The texts were divided into two groups with higher and lower holistic scores. The comparison between the two groups showed significantly more use of cohesive devices in the higher-level texts than the lower-level texts; meanwhile, the variety of cohesive devices used in the advanced level texts was more diverse. Moreover, strong correlation was found between coherence features and the holistic score of the texts. This conclusion supports Witte and Faigley's (1981) study on the important role of cohesion in enhancing writing quality. Moreover, a positive correlation between the use of cohesive devices and writing quality was also identified

by Liu and Braine (2005). Field and Oi (1992) and Norment (2002) also support the positive correlation between cohesion and proficiency level. In addition, studies that compared the difference between NS and NNS writing samples have confirmed the lack of cohesion in L2 writing (Crossley & McNamara, 2009).

However, there are also studies that failed to establish the significant correlation between cohesion and L2 proficiency level, such as Castro (2004) and Green (2012). Crossley and McNamara (2011) even found a negative correlation between the number of cohesive devices used in writing and proficiency levels.

Thus, in terms of cohesion and writing quality, many empirical studies have confirmed the critical role of performing cohesion in achieving higher writing quality. The differences between L1 and L2 writing in performing cohesion have also been established. Nevertheless, different even controversial findings have been found regarding the comparison between NNSs with various proficiency levels (Chen, 2008). One possible reason could be the different perspectives in defining the construct of cohesion or the different measuring approaches employed (Green, 2012). Table 9 presents the summary of the major studies mentioned above.

Table 9

Cohesion and Holistic Writing Quality

Study	Main Findings
Chen, 2008	- Cohesion in native and nonnative speakers' writing samples showed no difference.
Crossley & McNamara, 2009; Green, 2012	- Compared to native writing, L2 writing tends to lack cohesion.
Crossley & McNamara, 2011	- A negative correlation between the number of cohesive devices used in writing and proficiency levels.
Ferris, 1994	- Significantly more number of types of cohesive devices in the higher-level texts than the lower-level texts.
Field & Oi, 1992; Liu & Braine, 2005; Norment, 2002	- Positive correlation exists between the use of cohesive device and writing quality

Measurements

Manual and computational approaches are the two major approaches that have been used in measuring cohesion in written discourses. Manual approaches vary greatly according to the characteristics of the target texts. Meanwhile, the accuracy by manual approaches has been questioned due to the fallibility of hand counts and the subjective nature of intuitive judgement (Crossley & McNamara, 2009; Reid, 1992). Even though computational approaches are not perfect, due to the large size of corpora used in the present study, following Crossley and McNamara's (2009) methodology, Coh-Metrix 3.0 is employed as the computational tool.

The Coh-Metrix is a most recently developed software for computing linguistic indices that reveal cohesive properties of written and spoken texts (Graesser et al., 2004). Unlike the prior programs, the Coh-Metrix analyzes lexical, syntactic, and semantic properties of the texts that are related to cohesion. Meanwhile, the Coh-Metrix is built upon various existing resources and databases, including WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), the MRC Psycholinguistics Database (Coltheart, 1981), and the CELEX Database (Baayen, Piepenbrock, & van Rijn, 1993). The inclusive and comprehensive foundation allows the Coh-Metrix to process natural language and analyze linguistic features on various levels. More than 50 published studies have demonstrated the validity of the Coh-Metrix in detecting subtle differences in texts and discourses from genre sentence level (Crossley, Greenfield, & McNamara, 2008; Crossley, Louwerse, McCarthy & McNamara, 2007; Hall, Lewis, McCarthy, Lee, & McNamara, 2007; McCarthy, Briner, Rus, & McNamara, 2007). For instance, Crossley et al. (2008) conducted an exploratory study to examine the validity of Coh-Metrix in predicting readability at various levels of language, discourse, and conceptual analysis. The findings support the validity of Coh-Metrix variables (i.e., lexical, syntactic, and meaning construction

index) in accurately predicting cognitive reading processes ($r^2 = .86$). In examining texts in written format, Crossley and McNamara (2011) validated the indices from Coh-Metrix in distinguishing native and nonnative English writers' essays in terms of syntactic complexity, lexical diversity, and word frequency.

Summary

This section systematically introduces common measures for examining the lexical quality of a text. Three target lexical features are selected, namely lexical diversity, lexical sophistication, and cohesion, to represent the lexical quality of the writing samples. For lexical diversity, the MTLTD is employed as the quantitative computational tool; lexical sophistication is selected to represent the level of lexical richness and it is measured by the LFP. With respect to cohesion level, it is measured by the computational tool, Coh-Metrix 3.0.

Conclusion

The body of literature reviewed in this chapter highlights the critical role of lexical component in achieving higher L2 proficiency. In particular, decent L2 writing performance requires substantial vocabulary knowledge. Thus, researchers and educators focus on developing effective strategies to improve vocabulary researching and teaching. Based on naturally-occurred texts and advancements in various disciplines, Corpus Linguistics provides a reliable and valid approach to investigate language use, especially vocabulary usage, in the real world. In the fields of SLA and TESOL, Corpus Linguistics has been providing solid textual foundation and research tools to analyze how English is used by both native and nonnative speakers.

To better understand the performance and needs of NNSs, learner corpus research is derived from corpus studies in the fields of SLA and Applied Linguistics. Studies of learner corpus focus on the linguistic characteristics that L2 learners possess, which provide valuable

insights to pedagogical development. Moreover, the major methodological framework of learner corpus research, CIA, facilitates the exploration of the differences between L2 learners and NSs as well as across different groups of L2 learners.

Combining vocabulary research, Corpus Linguistics, and learner corpus research, empirical studies have been conducted to examine the lexical features of various groups of writers (e.g., native and nonnative speakers) in academic writing. Most studies report a more advanced use of lexical items by NSs. At the same time, NNSs often face difficulties in applying sufficient and appropriate types of vocabulary. However, research has largely focused on either one or only a limited number of aspects in lexical features. In addition, the comparison between native and nonnative speakers usually fails to involve various language groups to detect the potential influence of participants' L1. Therefore, to analyze lexical features of native and nonnative speakers' academic writing in a systematic and comprehensive way, the current study addresses three lexical features of both native and nonnative academic writing, including lexical diversity, lexical sophistication, and cohesion. The nonnative corpora consist of six language groups to enlarge the scope of L2 writing and investigate possible L1 influence on lexical performance. In the next chapter, methodology and detailed description of the procedures to conduct the study are introduced.

CHAPTER THREE: RESEARCH METHODOLOGY

Chapter Two presented the importance of developing ELs' essential writing skills to achieve academic success in English-medium institutions. The development of their vocabulary contributes to their holistic writing performance. Therefore, understanding the lexical features of ELs' academic writing can establish a foundation for developing pedagogies in vocabulary and writing instruction. Meanwhile, through the comparison between native and nonnative writing, educators can be informed of the specific differences that ELs have against their native speaking peers. Furthermore, comparing across NNSs from various language backgrounds provides insights of the diversity in TESOL.

Hence, the issues that the current study addresses are the differences of lexical features in academic writing 1) between native and nonnative English writers and 2) across all writers from various language backgrounds. To address the problem in a comprehensive manner, a corpus-based quantitative approach is employed to thoroughly examine three lexical features of the target population groups. The three lexical features are lexical diversity, lexical sophistication, and cohesion. By addressing the above issues, the present study aims to present representative profiles of both native and nonnative speakers' lexical features in academic writing. Thus, the gaps that ELs need to fulfill to achieve the language proficiency similar as their native-speaking peers are revealed. In addition, the comparison across different mother tongue groups sheds light on the diversity in nonnative English writers.

This chapter first restates the research questions and hypotheses of the study. An overview of the research design is followed. Next, a detailed description of the methods undertaken in the study is presented, including target population, selected corpora, sampling procedures, data collection procedures, and instrumentation.

Research Questions

1. Are there significant differences in lexical features between native and nonnative academic English writing, as measured by lexical diversity, lexical sophistication, and cohesion?
2. Are there significant differences in lexical features, as measured by lexical diversity, lexical sophistication, and cohesion, in academic English writing across all writers from various mother tongue backgrounds?

Hypotheses

Hypotheses for Research Question One

Lexical diversity

H₀: There are no significant differences in the level of lexical diversity (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of lexical diversity (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Lexical sophistication

H₀: There are no significant differences in the level of lexical sophistication (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of lexical sophistication (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Cohesion

H₀: There are no significant differences in the level of cohesion (as operationalized in this study) in academic writing between native and nonnative English writers.

H₁: Nonnative English writers' level of cohesion (as operationalized in this study) in academic writing is significantly lower than that of native English writers.

Hypotheses for Research Question Two

Lexical diversity

H₀: There are no significant differences in the levels of lexical diversity (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of lexical diversity (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Lexical sophistication

H₀: There are no significant differences in the levels of lexical sophistication (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of lexical sophistication (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Cohesion

H₀: There are no significant differences in the levels of cohesion (as operationalized in this study) in academic writing across all English writers from various mother tongue backgrounds.

H₁: The levels of cohesion (as operationalized in this study) in academic writing are significantly different across all English writers from various mother tongue backgrounds.

Orientation to Research Design

The present research follows a causal-comparative research design to determine whether the independent variable affects the outcome of the dependent variables by comparing multiple groups of individuals (Brewer & Kubn, 2012). A causal-comparative research design attempts to determine differences among variables without conducting actual manipulation of these variables. Thus, the research objectives of the current study meet the characteristics of a causal-comparative research design.

The independent variable (IV) in the present study is the writers' language backgrounds, which includes English and other six mother tongue backgrounds. The dependent variables (DVs) belong to three categories, including measures in lexical diversity, lexical sophistication, and cohesion. Because of different lengths between the texts, a covariate of text length was employed to strengthen the analyses and prevent skewness from text length. The data were derived from two corpora and quantitative measures were used to evaluate lexical quality of the texts. Both descriptive and inferential statistical methods of comparison were used to answer the research questions.

Two major corpora have been employed in the present study, namely the International Corpus of Learner English (ICLE) and the Louvain Corpus of Native Essay Writing (LOCNESS). The ICLE was employed to provide the learner English data. The second version of the ICLE (ICLE v2) consists of 6,085 texts and totals 3,753,030 words (Granger et al., 2009). Learner writing products from 16 mother tongue groups have been collected in the ICLE v2, including Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, and Tswana. To conduct the investigation in a comprehensive fashion at the same time control for the scope of the present study, six subcorpora were selected to represent learner English. The six subcorpora include Chinese, German, Japanese, Russian, Spanish, and Turkish. Detailed rationale for selection and sampling procedures are introduced shortly in the subsequent sections.

Adhering to the principle of conducting meaningful comparison, the LOCNESS has been selected as the main referential corpus, which contains 322 argumentative and literary compositions written by American and British native English-speaking university students. The total running words of the corpus is 324,304. Detailed introduction of this NS corpus is presented shortly. Table 10 illustrates the general information of the ICLE and the LOCNESS.

Table 10

The ICLE and the LOCNESS

	Writers' Language Backgrounds	Number of Essays	Number of Words
The ICLE	16 non-English backgrounds	6085	3,753,030
The LOCNESS	British and American NESs	322	324,304

For the statistical analyses, two Multivariate Analysis of Covariance (MANCOVA) have been implemented to answer the two major research questions of the differences in lexical

features in academic writing 1) between native and nonnative English writers and 2) across all writers from various language backgrounds. Figure 4 provides a visual overview of the research design that is further elaborated in subsequent sections.

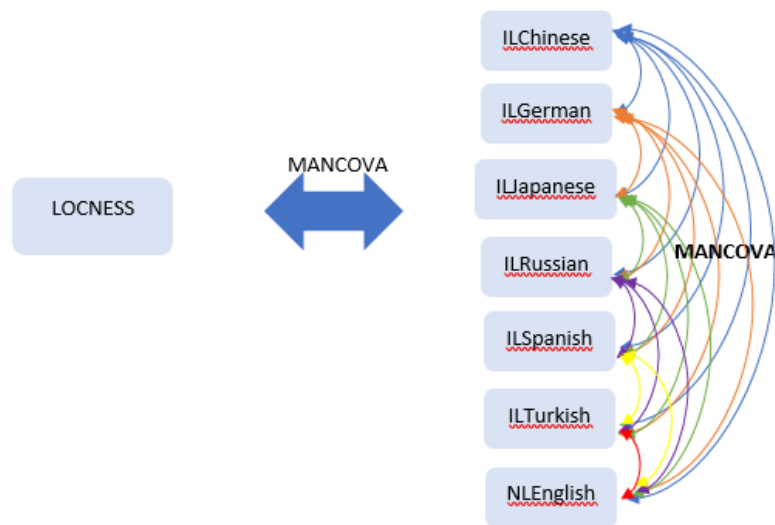


Figure 4: Research Design

Target Population and Selected Corpora

The target learner population of the study is advanced level ELs from universities in various countries. This group of learners is most likely to attend the same courses with NSs when they enter higher educational institutions in English-speaking countries. They are normally expected to write at the similar level as their native-speaking peers. Furthermore, advanced learners from universities in non-English speaking countries are usually at the similar age level as native speaking college students. Thus, the ICLE provides data from the desired population. The interlanguage data of this study are from six subcorpora of the ICLE, including the Chinese, German, Japanese, Russian, Spanish, and Turkish subcorpora.

The target native-speaking population is undergraduate students from British and American universities. Different from expert writers, native-speaking university students are also

at the developmental phase of academic writing. Thus, the present study does not suggest that NSs' writing is flawless. However, writing samples from NSs are more often used by instructors as the referential writing samples; additionally, they provide a realistic level for NNSs to compare with. After all, the ELs strive to be able to function in a regular academic class where their native-speaking counterparts are. For these reasons, the LOCNESS was selected as the native referential corpus for the present study. In this section, detailed introduction of the two major corpora and the rationale for data selection are presented.

The International Corpus of Learner English (ICLE): Nonnative English Speakers

In the early 1990s, academics started collecting foreign/second language learner data. In the early 2000s, the ICLE was made available to the academic community (Granger, Dagneaux, Meunier, & Paquot, 2002). The ICLE is a richly documented computer corpus, including authentic texts produced by foreign or L2 learners of English (Granger, 2003). This section introduces the development and design of the corpus, learner and task variables, size, and representativeness of the ICLE.

Corpus development and design

All learners included in the corpus were asked to complete a detailed profile questionnaire to provide more than 20 task and learner variables. For potential research purposes, there are shared features across all texts, which makes comparison across texts more reliable. Meanwhile, the texts also consist of individual features, such as different genders, mother tongues, and task settings. This characteristic enables the compilation of subcorpora to meet the systematic criteria.

The most distinguishing difference between the texts is the learners' various mother tongue backgrounds. This feature also directs the research design of the present study. In the first

version of the ICLE (Granger, 2002), 11 different language backgrounds were covered, including Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. Since the publication of the ICLE in 2002, the field of LCR has evolved greatly, and a wide array of research projects have employed the corpus as the basis for their analyses of interlanguages. Major journals in the field of TESOL, SLA, Corpus Linguistics, and Applied Linguistics have published empirical studies that used the first version of the ICLE as the database. Nesselhauf (2003) in *Applied Linguistics*, Flowerdew (2006) in *International Journal of Corpus Linguistics*, and Gilquin, Granger, and Paquot (2007) in *Journal of English for Academic Purposes* are a few examples.

In 2009, the second version of the ICLE (ICLE v2) was published with a higher amount and greater diversity of learner data as well as improved functionalities of the interface. Figure 5 shows the detailed task and learner variables that are recorded in the ICLE v2.

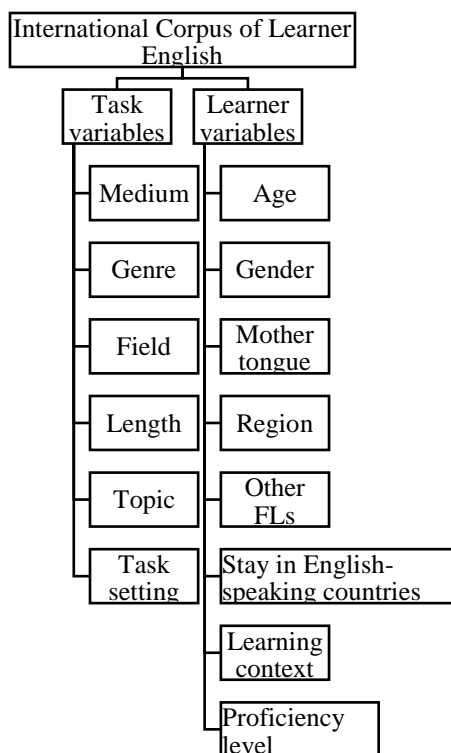


Figure 5: Task and Learner Variables of the ICLE

In addition to the basic database of texts, the ICLE v2 also contains a built-in concordance searching feature which allows simple and complex searches (Granger et al., 2009). All textual data were lemmatized and part-of-speech tagged with the Constituent Likelihood Automatic Word-tagging System (CLAWS) C7. Therefore, using the concordance searching engine provides the word form of the word unit and gives part-of-speech (POS) tag and its lemma.

With more comprehensive and detailed learner profile information, feedback of the first version from professionals, and advancement of the searching engine, the ICLE v2 is considered as a much more robust and reliable learner corpus than the first version. Various studies have been conducted based on data from the ICLE v2 and published in major journals in TESOL, SLA, Corpus Linguistics, and Applied Linguistics. Durrant and Schmitt (2009) in *International Review of Applied Linguistics in Language Teaching*, Crossley and McNamara (2009) in *Journal*

of *Second Language Writing*, and Thewissen (2013) in *The Modern Language Journal* are a few examples. The examination of the corpus in those studies further justified the reliability of the corpus. Thus, the present study employs the ICLE v2 as the foundational corpus to examine lexical features of interlanguages. Next, I introduce the general task and learner variables of the corpus in detail.

Learner and task variables

The learners in the ICLE are all young adults (university undergraduates) with higher intermediated to advanced proficiency level of English. English is considered as the foreign language rather than the second language for the learners. The collected texts are all academic writing from the learners with 91% of those are argumentative essays. The argumentative essays are considered as an appropriate genre type to investigate discourse-oriented lexical and grammatical features (Biber & Gray, 2013). In terms of text length, the length of each writing sample ranges from 384 words (Tswana) to 893 words (Dutch), with the average text length being 617 words. Regarding the topics selected for the texts, all subcorpora follow the same list of suggested topics provided by the leading team. Table 11 shows the 10 most popular topics and the subcorpora that has the highest proportion of the specific topic (Granger et al., 2009). Lastly, 62% of the essays were written in an untimed setting and 61% were not written under exam conditions; also, 48% were written with the support of reference tools. The overview of the task variables is presented in Table 12 (Granger et al., 2009).

Table 11

Top 10 Topics in the ICLE

Essay Topic	Number of Essays	Country of Origin
Some people say that in our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?	491	29% Bulgarian
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	249	22% Turkish
Poverty is the cause of the HIV/AIDS epidemic in Africa.	243	100% Tswana
Marx once said that religion was the opium of the masses. If he was alive at the end of the 20 th century, he would replace religion with television.	237	19% Russian
The prison system is outdated. No civilized country should publish its criminals: it should rehabilitate them.	176	32% Tswana
Discuss the advantages and disadvantages of banning smoking in restaurants.	156	100% Chinese
Discuss the advantages and disadvantages of using credit cards.	149	100% Chinese
Feminists have done more than harm to the cause of women than good.	139	23% Russian
In the words of the old song “Money is the root of all evil”.	133	22% Russian
In his novel “Animal Farm”, George Orwell wrote “All men are equal: but some are more than others”. How true is this today?	127	39% Bulgarian

Source: Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). *International corpus of learner English V2*. Presses Universitaires de Louvain.

Table 12

Task Variables in the ICLE

Medium	Genre	Field	Text Length	Topic	Task Setting
English	Argumentative essays 91%	Vary	617 in average	Vary	62% timed setting; 61% non-exam condition; 48% written with reference tools

Source: Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). *International corpus of learner English V2*. Presses Universitaires de Louvain.

For learner variables, almost all learners were young adults with average age of 22.30 and female learners were the majority across all subcorpora (76%). In terms of the mother tongue

background, the ICLE v2 included five more different language backgrounds than did the first version. Thus, there are 16 different mother tongue backgrounds, including Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana. Other languages that were spoken at home were also recorded as the metadata of the corpus. The region variable is included for specifying languages that are spoken in more than one country. For instance, Dutch is spoken in both Belgium and the Netherlands; German is spoken in Germany, Austria, and Switzerland. Knowledge of other foreign languages and the time spent in an English-speaking country were also recorded for distinguishing learner backgrounds.

Two fuzzy variables include learning context and proficiency. In terms of the learning context, it can be difficult to distinguish EFL and ESL settings. For instance, some scholars argue that English can be considered as the L2 in Hong Kong. However, this opinion is not held by many other professionals. The most certain point is that all learners represented in the corpus have learned English primarily in a classroom setting.

The other fuzzy variable is the English proficiency level of the writers. Even though the corpus intended to collect written texts from university students with advanced proficiency of English, more detailed inspection of some of the texts reveals clear differences in writing quality. Twenty essays from each of the 16 subcorpora were rated based on the Common European Framework of Reference for Languages (CEFR). The CEFR provides a six-level scale to assess foreign language proficiency. From low to high proficiency level, the six levels include A1 and A2, B1 and B2, C1 and C2. From level A to C, the three broad levels are described as Basic User, Independent User, and Proficient User (“The CEFR Levels,” 2018). It can be seen from Table 13 that the ratings of the 20 selected essays in each subcorpora vary from B2 to C2.

Table 13

CEFR Results - 20 Essays Per Subcorpus

Mother Tongue	B2	C1	C2	Total
Bulgarian	2	16	2	20
Chinese	19	1	0	20
Czech	11	9	0	20
Dutch	1	11	8	20
Finnish	3	8	9	20
German	1	12	7	20
Italian	10	9	1	20
Japanese	18	2	0	20
Norwegian	8	7	5	20
Polish	1	12	7	20
Russian	3	15	2	20
Spanish	12	8	0	20
Swedish	0	14	6	20
Tswana	18	0	2	20
Turkish	16	4	0	20
Total	126	139	55	320

Source: Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). *International corpus of learner English V2*. Presses Universitaires de Louvain.

Corpus size and representativeness

The number of essays included in the ICLE v2 is 6,085 with 3,743,030 words in total. For each native language subcorpus, approximately 200,000 tokens are included. Only the subcorpus of Chinese writers has more than 490,000 running words. Table 14 illustrates the detailed statistics of the number of essays and words in each subcorpus.

Table 14

Subcorpora Size in the ICLE

Native Language Subcorpus	Number of Essays	Number of Words
Bulgarian	302	200,194
Chinese	982	490,617
Czech	243	201,687
Dutch	263	234,723
Finnish	390	274,628
French	347	226,922
German	437	229,698
Italian	392	224,222
Japanese	366	198,241
Norwegian	317	211,725
Polish	365	233,920
Russian	276	229,584
Spanish	251	198,131
Swedish	355	200,033
Turkish	280	199,532
Tswana	519	199,173
ICLE v2	6,085	3,753,030

Source: Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). *International corpus of learner English V2*. Presses Universitaires de Louvain.

Six selected mother tongue backgrounds

Six subcorpora were chosen in the present study from the ICLE v2 to represent learner interlanguages, including mother tongues of Chinese, German, Japanese, Russian, Spanish, and Turkish. Statistics from the United Nations Educational, Scientific and Cultural Organization (UNESCO, 2016) indicate the global flow of international students on tertiary-level of education. Choosing the major English-speaking countries (e.g. the US, the UK, Australia, New Zealand) as the destinations, these six mother tongues represent the top countries of origin for international students.

As stated in Chapter Two, the framework of CIA requires a reference corpus to conduct the systematic comparison between interlanguages and the referential textual data. The referential corpus can be a corpus of native speaker or expert textual data (Granger, 2015). The

diversity of English varieties around the global is acknowledged, so it is certainly not a simple task to establish one or two types of English as the so-called “norm”. However, the ultimate purpose of the current study is to provide pedagogical insights for instructing academic English writing to learners who plan to achieve higher academic performance in English-speaking educational settings. With the current situation in academia, the learners are often expected to perform at least at a similar level as their native-speaking peers. For the L2 learners, reaching an expert level of writing can be considered as an eventual objective rather than a realistic goal. Therefore, for the specific EL population in the current study (i.e., young university adults), it is more reasonable to compare their writing to that of similar grade level NSs rather than experts. In addition, it is crucial to emphasize that the selected NS corpus in the current study does not represent the perfect academic writing quality by any means. Rather, it is used as a reference to examine the potential lexical differences between native and nonnative speakers’ academic writing. Hence, the LOCNESS is chosen as the referential corpus. In the next section, this corpus is introduced in detail.

The Louvain Corpus of Native Essay Writing (LOCNESS): Native English Speakers

The project of the LOCNESS was led by Silva Granger and her colleagues at the Université Catholique de Louvain. The corpus was compiled with the intention to create a parallel NS corpus of the ICLE. The LOCNESS is made up of British pupils’ A level essays, British university students’ essays, and American university students’ essays. The total number of running words is 324,304. Table 15 illustrates the distribution of running words across the three subcorpora of the LOCNESS.

Table 15

Subcorpora Size in the LOCNESS

British pupils' A level essays	British university students' essays	American university students' essays
60,209	95,695	168,400

The genres of essays collected in the LOCNESS are mostly argumentative and literary essays. Majority of the native speaking writers' age varied from 18 to 23. The essays contain a wide range of topics, such as animal testing, nuclear power, water pollution, and so forth. The writing settings differ across each subcorpora, both timed and untimed settings can be found.

Table 16 presents the general components and distribution of the LOCNESS.

Table 16

Essays in the LOCNESS

British Essays: University students		American Argumentative Essays	American Literary-Mixed Essays
Year collected: 1991	Year collected	1995	1995
Timed essays: 57 (exam) Untimed essays: 33 (not exams)	Universities	Marquette University; Indiana University at Indianapolis; Presbyterian College, South Carolina; University of South Carolina; University of Michigan. (5)	Presbyterian College, South Carolina. (1)
	Example topics	- Women in combat - Curfew - Abortion - Rules and regulations	- An aspect of studying ethnic American Literature - Who is Hamlet?
Genre: - literary (e.g. "French Intellectual tradition"): 39 - expository/historical (e.g. "French higher education"): 18 - argumentative essays: 33 - other (A levels): 60,209 words in total	Total number of essays	176	56
	Times essays	88	56
	Untimed essays	88	0
	Age	17-48; most 18-22	18-25
	Average words per essay	850	336
Average words per essay: - literary: 1501 - expository/historical: 1007 - argumentative: 576			
Total words: 155,904	Total words	149,574	18,826

As the ICLE was employed in multiple empirical studies, the LOCNESS has also been used in various published articles in major journals and books in the field of TESOL, SLA, Corpus Linguistics, and Applied Linguistics. Examples can be found in Granger and Tyson (1996) in *World Englishes*, De Cock (2000) in the book *Corpus Linguistics and Linguistic Theory*, Altenberg and Granger (2001) in *Applied Linguistics*, Aijmer (2002) in the book *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Durrant and Schmitt (2009) in *International Review of Applied Linguistics in Language Teaching*, and Laufer and Waldman (2011) in *Language Learning*.

In sum, there are two major corpora selected to provide the foundational data for analyzing research questions in the present study. On one hand, the ICLE provides the data of learner English (interlanguages). On the other hand, the LOCNESS was chosen to function as the referential corpus, providing NS data. Table 17 presents the published studies that have employed the above two corpora for analyses.

Table 17

Publications Based on the ICLE and the LOCNESS

	Author(s)	Year	Title	Journal/Book
The ICLE	Nesselhauf	2003	The use of collocations by advanced learners of English and some implications for teaching	<i>Applied Linguistics</i>
	Flowerdew	2006	Use of signalling nouns in a learner corpus	<i>International Journal of Corpus Linguistics</i>
	Gilquin, Granger, & Paquot	2007	Learner corpora: The missing link in EAP pedagogy	<i>Journal of English for Academic Purposes</i>
The ICLE v2	Durrant & Schmitt	2009	To what extent do native and non-native writers make use of collocations?	<i>International Review of Applied Linguistics in Language Teaching</i>
	Crossley & McNamara	2009	Computational assessment of lexical differences in L1 and L2 writing	<i>Journal of Second Language Writing</i>
	Thewissen	2013	Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus	<i>The Modern Language Journal</i>
The LOCNESS	Granger & Tyson	1996	Connector usage in the English essay writing of native and non-native EFL speakers of English	<i>World Englishes</i>
	De Cock	2000	Repetitive phrasal chunkiness and advanced EFL speech and writing	<i>Corpus Linguistics and Linguistic Theory</i>
	Altenberg & Granger	2001	The grammatical and lexical patterning of MAKE in native and non-native student writing	<i>Applied Linguistics</i>
	Aijmer	2002	Modality in advanced Swedish learners' written interlanguage	<i>Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching</i>
	Durrant & Schmitt	2009	To what extent do native and non-native writers make use of collocations?	<i>International Review of Applied Linguistics in Language Teaching</i>
	Laufer & Waldman	2011	Verb-noun collocations in second language writing: A corpus analysis of learners' English	<i>Language Learning</i>

Sample Size and Sampling Procedures

Based on the ICLE and the LOCNESS, the present study used purposive sampling method to select nonnative and native written texts. The purposive sampling method was employed to construct parallel and comparable corpora for comparisons. As stated previously, six subcorpora from the ICLE have been selected to represent six non-English mother tongue backgrounds. The genre selected across all subcorpora was argumentative essays. As a result, 100 argumentative essays were randomly selected from each corpus, which yielded 700 argumentative essays (424,363 tokens) in total for the present study. Procedures on how the study arrived at these figures are introduced in depth in the following section.

Sample Size Determinations

The number of independent and dependent variables needed to be determined before deciding the sample size of the study. The current study included seven language groups (six non-English, one English), which was the IV. However, the classification and calculation of the DVs were more complex. There were three categories of the major lexical features. Table 18 shows the measurements and the categories to which they belong. Table 19 elaborates the detailed indices of the selected cohesion constructions.

Table 18

Summary of the Target Measurements of Lexical Features

	Lexical Diversity	Lexical Sophistication	Cohesion
Measurements	MTLD	The LFP: coverage of 1 st 1000 words, 2 nd 1000 words, AVL, off-list words. AVL	Referential cohesion, LSA, connectives (see Table 19)

Regarding the measurement of cohesion, Coh-Metrix has been selected as the computational tool to evaluate the cohesion level of all writing samples. Building upon the study of Crossley and McNamara (2009), 19 indices have been carefully chosen to represent the aspects of referential cohesion, latent semantic analysis (LSA), and connectives (Table 19).

Referential cohesion, or coreference, refers to overlap in content words, or word repetition, between consecutive and adjacent sentences as well as between all of the sentences in a paragraph or text (McNamara, Graesser, McCarthy, & Cai, 2014). Different types of coreference are measured in Coh-Metrix, including noun overlap, argument overlap, stem overlap, and content word overlap.

Noun overlap measures the proportion of sentences with overlapping nouns in a text. No deviation is allowed in the morphological forms of the nouns. For instance, the word *university* only has one overlapping noun *university*; the plural form *universities* is not considered as an overlap.

Argument overlap considers overlap between the head nouns and pronouns. For instance, the prior example of *university* can have an argument overlap *universities* because they share the same head noun. An instance of overlap between pronouns can be *he* and *he*. Here, the term “argument” refers to noun/pronoun arguments which are contrasted with verb/adjective predicates (Kintsch & Van Dijk, 1978). The argument overlap is less strict with the morphological form when compared to noun overlap.

To illustrate the concept above with precise examples, the following few sentences were extracted from a Chinese writer’s essay from the corpus of the current study:

Not only one or two many of my peers view our *university degrees* with scepticism. They often said: “what *we* learn in *universities* are mainly from books which usually are

theoretical rather than practical. After years of study, *we* graduate with *degrees*. But *we* got nothing more than theoretical knowledge.

In this example, the words *university* and *universities* do not construct the noun overlap due to different morphological form. However, the word *degrees* appears in both the first sentence and the third sentence; thus, this is an example of noun overlap. In terms of argument overlap, the pronoun *we* appears for multiple times, which forms the argument overlap.

Stem overlap considers overlap between a noun in one sentence and a content word in another sentence. Content words can be nouns, verbs, adjectives, and adverbs. The content word in the other sentence must share a common lemma with the noun. For instance, the noun *price* can have a stem overlap in another sentence with the word *priced*.

Lastly, content word overlap refers to the proportion of explicit content words (i.e. nouns, verbs, adjectives, and adverbs) that are shared between sentences. For two pairs of sentences with the same sentence length, the pair with more content words overlap has a higher proportion. The following two pairs of adjacent sentences are from a Japanese writer's essay and a British writer's essay respectively:

Pair 1 (Japanese):

But about ownership of *land* they cannot compromise.

They cannot leave off familiar *land* where they have been living for a long time.

Pair 2 (British):

Whether Britain will lose its *sovereignty* or not, is entirely a personal viewpoint.

But will stepping out of the single market be worthwhile preserving our *sovereignty*?

Here, both pairs have one content word (noun) overlap: *land* and *sovereignty*. However, the text length of the British writer's essay is longer than that of the Japanese writer's; therefore,

the proportion of content word overlap of the British writer's writing is lower than that of the Japanese writer's in this case.

In sum, referential cohesion considers the overlap between sentences and in a text from a word level. Eight indices were selected to represent the coreference of noun overlap, stem overlap, argument overlap, and content word overlap. For each type of coreference, both overlap between adjacent sentences and all sentences in a text are measured. In order to keep the further statistical analyses consistent, only means of the indices were selected. Table 19 shows the summary of the indices and their descriptions.

In terms of the LSA, it provides the measures of semantic overlap between sentences or between paragraphs. The semantic overlap includes the overlap between explicit words and words that implicitly similar or related in meaning. There are four types of LSA indices included in Coh-Metrix 3.0, including LSA similarities between adjacent sentences, all sentences in a paragraph, adjacent paragraphs, and LSA given/new score.

Different from referential cohesion, LSA overlap and similarities consider the cohesion at the level of not only word forms but also word meanings. For instance, if one sentence has the word *driver*, a relatively high degree of semantic overlap can be found with words such as *car*, *street*, *road*, and so forth. The first three types of LSA indices are comparatively easy to understand with the only difference of the scope. The fourth type, LSA give/new score, measures how much given versus new information exists in each sentence in a text, compared with the content of prior information. The ratio can be understood as $G/(N+G)$. When there is more given information in a text and less new information, the ratio approaches 1, which represents a higher level of cohesion. In contrast, if there is less given information and more new information, the ratio score approaches 0, which indicates a lower level of cohesion. To illustrate the concept of

LSA given/new score, I randomly selected a sentence from the NS corpus in this study as an example.

There is no doubt an integrated *market* would have multiple benefits for the countries involved. *Businesses* and other *trading* organizations are preparing themselves for 1992, and the *single market*. Trade will be easier, with no frontier controls allowing free flowing transportation of merchandise.

In this example, *market* is a new word when it first mentioned, while *business*, *trading organizations*, and *single market* are all coreferential with it. Hence, the latter words and phrases are given information even though there are lexical differences that have to be bridged inferentially.

Again, for the consistency in further statistical analyses, only mean scores of the four types of LSA indices were selected. The four indices and their descriptions are shown in Table 19.

The last measurement of cohesion is connectives. There are incidence scores for all connectives and six individual types of connectives. Descriptions and examples of different types of connectives can be found in Table 19.

Table 19

Measurements of Cohesion

Category	Index	Description
Referential cohesion	28 CRFNO1	Noun overlap, adjacent sentences, binary, mean
	29 CRFAO1	Argument overlap, adjacent sentences, binary, mean
	30 CRFSO1	Stem overlap, adjacent sentences, binary, mean
	31 CRFNOa	Noun overlap, all sentences, binary, mean
	32 CRFAOa	Argument overlap, all sentences, binary, mean
	33 CRFSOa	Stem overlap, all sentences, binary, mean
	34 CRFCWO1	Content word overlap, adjacent sentences, proportional, mean
	36 CRFCWOa	Content word overlap, all sentences, proportional, mean
LSA	38 LSASS1	LSA overlap, adjacent sentences, mean
	40 LSASSp	LSA overlap, all sentences in paragraph, mean
	42 LSAPP	LSA overlap, adjacent paragraphs, mean
	44 LSAGN	LSA given/new, sentences, mean
Connectives	50 CNCAI1	All connectives incidence
	51 CNCCaus	Causal connectives incidence, e.g., <i>because, so</i>
	52 CNCLogic	Logical connectives incidence, e.g., <i>and, or</i>
	53 CNCADC	Adversative and contrastive connectives incidence, e.g., <i>although, whereas</i>
	54 CNCTemp	Temporal connectives incidence, e.g., <i>first, until</i>
	55 CNCTempx	Expanded temporal connectives incidence, e.g., <i>finally, last week</i>
	56 CNCAdd	Additive connectives incidence, e.g., <i>and, moreover</i>

In conclusion, the three measurements of cohesion selected in the current study include 1) referential cohesion, which considers the cohesion at the level of word forms; 2) LSA, which considers the cohesion at the semantic level; and 3) connectives, which sums up the use of different types of connectives.

Nevertheless, it is unrealistic and unpractical to conduct further statistical analyses with all of the selected indices for the measurement of cohesion. Thus, an Exploratory Factor Analysis (EFA) has been conducted to deduct the indices of cohesion in order to reach a manageable number of DVs.

Exploratory Factor Analysis (EFA)

Exploratory factor analysis (EFA) is a statistical approach which explores the latent variables behind a set of variables or measures (Grant & Fabrigar, 2007). Through exploring the underlying structure of correlations among observed variables, the major goal of EFA is to specify a small number of factors that can account for the correlations among a set of measured variables. In other words, an EFA is conducted to compress the existing large number of variables into a few manageable variables that can still represent the whole construct. For the present study, the purpose of conducting the EFA is to reduce the 19 indices (see Table 19) that measure cohesion to a smaller number of factors in order to make them easier to manage as the DVs in statistical analyses.

A principal axis factor analysis, which is an EFA approach, was conducted on the 19 items with varimax rotation. The purpose of principal axis factor analysis is to obtain parsimonious representation of observed correlations between variables by latent factors. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, $KMO = .743$, which is a meritorious sample size according to Hushon and Sofroniou (1999). An initial

analysis was run to obtain Eigenvalues for each factor in the data. Six factors had Eigenvalues over Kaiser's criterion of 1 and in combination explained about 83% of the variance. However, to control the analysis in an attainable scope for the study and based on the original classification of these 19 indices in Coh-Metrix (See Table 19), I restricted the number of factors to three for another analysis. The three fixed factors still explained more than 64% of the variance. In addition, the three new factors were in compliance with the original classification in Coh-Metrix.

The items clustered in the same factor suggest that Factor 1 represents the indices related to referential cohesion. Indices clustered in Factor 2 represent LSA and Factor 3 includes indices that are mostly related to connectives.

Table 20 concludes the three factors extracted from the EFA. It can be seen that these three constructs and the including indices from the EFA are identical with the original classification in Table 19. In short, the EFA determined the three DVs that can measure cohesion. These three variables were referential cohesion, LSA, and connectives.

Table 20

Three Factors Extracted from the EFA

Factors	Indices included	Description
Factor 1: Referential Cohesion	28 CRFNO1	Noun overlap, adjacent sentences, binary, mean
	29 CRFAO1	Argument overlap, adjacent sentences, binary, mean
	30 CRFSO1	Stem overlap, adjacent sentences, binary, mean
	31 CRFNOa	Noun overlap, all sentences, binary, mean
	32 CRFAOa	Argument overlap, all sentences, binary, mean
	33 CRFSOa	Stem overlap, all sentences, binary, mean
	34 CRFCWO1	Content word overlap, adjacent sentences, proportional, mean
	36 CRFCWOa	Content word overlap, all sentences, proportional, mean
Factor 2: LSA	38 LSASS1	LSA overlap, adjacent sentences, mean
	40 LSASSp	LSA overlap, all sentences in paragraph, mean
	42 LSAPP1	LSA overlap, adjacent paragraphs, mean
	44 LSAGN	LSA given/new, sentences, mean
Factor 3: Connectives	50 CNCA11	All connectives incidence
	51 CNCCaus	Causal connectives incidence
	52 CNCLogic	Logical connectives incidence
	53 CNCADC	Adversative and contrastive connectives incidence
	54 CNCTemp	Temporal connectives incidence
	55 CNCTempx	Expanded temporal connectives incidence
	56 CNCAdd	Additive connectives incidence

Sample size estimation

After determining IV and DVs, two MANCOVA tests have been carried out to compare the levels of lexical diversity, lexical sophistication, and cohesion in academic writing between NNEs and NESs as well as across all seven language groups. Table 21 elucidates the two MANCOVA tests.

Table 21

Two Types of MANCOVA

Analyses	Type of Comparison	Lexical Features (DVs)	Covariate
MANCOVA 1	NNESs vs. NESs	- Lexical diversity - Lexical sophistication	Text length
MANCOVA 2	Across seven groups	(coverages of 1 st 1,000, 2 nd 1,000, AWL, off-list, AVL words) - Cohesion	

Among the two MANCOVA tests, the second MANCOVA test contained the highest number of IV and DVs. There was one IV (i.e., language background) which included seven levels, namely Chinese, German, Japanese, Russian, Spanish, Turkish, and native English writers. The DVs included nine lexical features, which belonged to three major categories, namely lexical diversity, lexical sophistication, and cohesion. Therefore, the sample size for the present study was determined based on the requirement of the second MANCOVA test.

Stevens (2009) provided a useful table for the calculation of MANOVA sample sizes. The table shows that at an alpha of .05 and power of .8 to detect a moderate effect size with a MANOVA consisting of four groups and six DVs, at least 74 cases are needed in each group. If five groups are involved and other factors remain the same, 82 cases are needed in each group. When there are six groups, at least 90 cases are required for each group. The current study consists of seven language groups (six non-English, one English) and nine DVs are measured. Thus, one hundred cases per group can be a reasonable estimation for achieving a moderate effect size with the conventional 80% power at a .05 level of significance.

In addition, a confirmative power analysis for a MANOVA with seven levels and nine DVs was conducted in G*Power to verify the estimated sample size (Faul, Erdfelder, Buchner, & Lang, 2014). Using the preliminary value of Pillai V = .4, seven number of groups and nine

response variables, the effect size f^2 (V) was calculate as .07. With an alpha of .01, a power of .90, and the calculated effect size ($f^2 = .07$), the desired sample size is 126. Thus, the estimated sample size of 100 per group, 700 in total, was validated and proved to be more than enough.

Sampling Procedures

The present study employs existing corpora to obtain data for analyses. The research does not involve collecting data through intervention nor interaction with the individual; moreover, no identifiable private information is included in the existing corpus data. Thus, the current research does not include human subjects (“Human Subject Regulation,” 2016). In this case, the process of sampling and data collection belongs to the “not human subject research” category of the Institutional Review Board (IRB) review process (see Appendix for a copy of the IRB Approval and Explanation of Research). Upon approval from the IRB of the University of Central Florida, data collection procedures occurred in the following steps:

1. Purchasing the International Corpus of Learner English v2 (Handbook + CD-Rom) from i6doc.com.
2. Compiling six nonnative speaker argumentative-essay corpora from the ICLE, including the mother tongues of Chinese, German, Japanese, Russian, Spanish, and Turkish.
3. Randomly selecting 100 essays from each native language subcorpus.
4. Compiling native speaker corpus of argumentative essays from the LOCNESS.
5. Randomly selecting 100 essays from the native speaker corpus.
6. Preparing essays to text format for analyses.

Data Extraction and Measurements

Three lexical features, including lexical diversity, lexical sophistication, and cohesion, of native and nonnative English speakers’ academic writing samples have been analyzed in the

present study. The rationale and detailed descriptions of the selected lexical features can be found in Chapter Two. Here, the instrumentation employed to conduct the measurements is presented. Table 22 concludes this section by summarizing the instrumentation for measuring the lexical features and the corresponding data types of each lexical feature.

Lexical Diversity

Various empirical studies have demonstrated the positive contribution of lexical diversity to the holistic quality of academic writing. Moreover, the potential differences between native and nonnative speakers' writing in lexical diversity are supported by empirical studies (Laufer & Nation, 1995; Yu, 2009). Here, after comparing several approaches that measure lexical diversity from the literature (see Chapter Two), theoretical explanations and empirical studies have supported the reliability of the MTLTD as a robust measurement for lexical diversity without being influenced by text length (McCarthy & Jarvis, 2010). Thus, in the current study, the MTLTD value was employed for measuring the lexical diversity levels of the writing samples. Chapter Two introduced the detailed operationalization of the MTLTD.

As stated in Chapter Two, the Coh-Metrix 3.0, developed by Graesser et al. (2004), has been tested as a reliable tool for analyzing lexical features of a text. One of the 106 linguistic indices in Coh-Metrix 3.0 provides the MTLTD value to represent the level of lexical diversity. The Coh-Metrix 3.0 can be freely accessed online (<http://cohmetrix.com/>), which allows user to copy and paste essays in text format into the program for analysis.

Lexical Sophistication

Lexical sophistication is considered as one of the key components of lexical richness. Employing a higher percentage of less frequent vocabulary often indicates higher levels of language proficiency and better quality of writing (Laurfer, 1994; Laufer & Nation, 1995; Silva,

1993). Laufer (1994) and Laufer and Nation (1995) proposed the LFP as a comprehensive method to measure the level of lexical sophistication of a text. The LFP includes the coverages of the first 1,000 words, second 1,000 words, academic words, and off-list words in a text.

To date, the LFP has been employed as the main approach for measuring lexical sophistication. Despite some existing questions and criticisms, the LFP has been validated by various empirical studies, which indicates its ability to demonstrate the stair-step lexical profiles of academic writing. The program is currently freely accessible on the Compleat Lexical Tutor website (Cobb, n.d.).

Some scholars have questioned the appropriateness of the AWL in representing the usage of academic words (Hyland & Tse, 2007; Nation & Webb, 2010; Paquot, 2010; Schmitt, 2010;). To obtain a more reliable result, the AVL (Gardner & Davies, 2013) is used as another referential word list to examine the coverage of academic vocabulary in the texts to complement the AWL coverage. A detailed introduction of the AVL was presented in Chapter Two. The AVL is currently available as an integral part of an online resource found at www.wordandphrase.info/academic.

Cohesion

Empirical studies support the critical role of cohesion in connecting ideas in a composition to help readers comprehend the content with less disruption (Ferris, 1994; Field & Oi, 1992; Norment, 2002; Witte & Faigley, 1981). Meanwhile, extant literature has also demonstrated that more diverse cohesion devices used in a text help enhance the cohesion level and holistic quality of the text (Ferris, 1994). In addition, Crossley and McNamara (2009) note the difference in the level of cohesion between native and nonnative speakers' writing. NNSs were found to lack cohesion in L2 English writing.

Prior research studies have employed cohesive devices as the major representation of the level of cohesion in a text. Both manual and computational approaches have been used in counting cohesive devices and measuring the level of cohesion. Nevertheless, scholars have pointed out the fallibility of hand counts and the subjective nature of intuitive judgement (Crossley & McNamara, 2009; Reid, 1992). Thus, with more advanced techniques, computational approaches can work more efficiently, effectively, and accurately in analyzing large size corpora.

Coh-Metrix 3.0, developed by Graesser, Crossley, McNamara and their colleagues, has been validated by empirical research in analyzing cohesion from a mathematic and quantitative perspective (Graesser et al., 2004). The current study employed Coh-Metrix 3.0 to reveal the cohesion of the compositions. After conducting the EFA as stated in prior section, three constructs (i.e., referential cohesion, LSA, and connectives) have been generated to represent the level of cohesion.

Lastly, a summary of the instrumentation used in measuring the three lexical features is presented in Table 22. The target values and data types of the lexical measures are also included.

Table 22

Summary of the Dependent Variables

	Lexical Diversity	Lexical Sophistication	Cohesion
Instrumentation	Coh-Metrix	Lexical Frequency Profiles (LFP)	Coh-Metrix
Value	MTLD	Coverages of 1 st 1000, 2 nd 1000, AVL, off-list words. Supplemented by coverage of AVL	Referential cohesion, LSA, connectives
Data Type	Continuous	Continuous	Continuous

Conclusion

This chapter described the research design, target population and corpora, sampling and data collection procedures, and instrumentation. The research design was based on the updated framework of the CIA, CIA² (Granger, 2015). Built upon the methodological framework, comparisons of lexical features have been made between native and nonnative English writers as well as across all writers from seven language backgrounds.

The target population of NNSs is advanced ELs in non-English-speaking countries; the referential native speaking population is native English-speaking university students. The learner English corpora are six subcorpora selected from the ICLE, including the mother tongues of Chinese, German, Japanese, Russian, Spanish, and Turkish. The NS corpus is the LOCNESS, which includes essays written by both British and American college undergraduate students. For the seven selected subcorpora, 100 argumentative essays were randomly selected from each subcorpus. A total of 700 texts have been analyzed in the study. The total tokens are 424,363 words.

Lastly, various computational tools were employed to measure the lexical features. Lexical diversity has been measured by the construct of the MTLT in Coh-Metrix 3.0; lexical sophistication has been examined by the LFP and complemented by the AVL coverage; finally, cohesion has been investigated through three constructs generated from indices in the Coh-Metrix 3.0.

Chapter Four discusses the results of the comparisons and reveals the responses to the major research questions. Following the findings, Chapter Five addresses the pedagogical implications of the study as well as future research directions that could follow up on the results derived in this study.

CHAPTER FOUR: RESULTS

This chapter presents the findings of the present study which investigated the differences in lexical features 1) between native and nonnative English writers' academic writing and 2) across essays from all writers with various language backgrounds. The chapter revisits the research questions, associated hypotheses, and research design previously addressed in Chapter Three. The descriptive statistics of the sample are then elucidated. The chapter proceeds to describe the data screening, normality, and assumption checks conducted prior to data analysis. The results from the two MANCOVAs are included in the concluding section.

Research Questions

The study was designed to thoroughly examine the lexical features of the selected academic writing samples and compare the differences 1) between native and nonnative English writers and 2) across all writers from seven different language backgrounds. To achieve this objective, two research questions guided the analyses. These questions are presented below along with their corresponding hypotheses.

1. Are there significant differences in lexical features between native and nonnative academic English writing, as measured by lexical diversity, lexical sophistication, and cohesion?

As suggested by previous studies which revealed remarkable differences between native and nonnative writers in vocabulary (Crossley & McNamara, 2009; Ferris, 1994; Field & Oi, 1992; Flowerdew, 2010; Grant & Ginther, 2000; Reid, 1992), it was hypothesized that NNSs' levels of lexical diversity, lexical sophistication, and cohesion (as operationalized in this study) in academic writing would be significantly lower than those of native English writers.

2. Are there significant differences in lexical features, as measured by lexical diversity, lexical sophistication, and cohesion, in academic English writing across all writers from various mother tongue backgrounds?

Based on evidence from existing literature regarding the diversity of NNSs in terms of academic writing (Díaz-Bedmar & Papp, 2008; Hong & Cao, 2014; Paquot, 2008, 2010), it was posited that there would be significant differences in lexical diversity and lexical sophistication in academic writing between writers from various language backgrounds.

Relatively few studies have examined the difference in cohesion between L2 writings from various language backgrounds. Hong and Cao (2014) addressed the intergroup homogeneity and heterogeneity in L2 writing. However, given the diverse mother tongue backgrounds in the current study, it was hypothesized that there would be significant differences in cohesion between at least two groups of nonnative writers.

To answer these research questions and test the directional and nondirectional hypotheses, the study followed a quantitative research design. The statistical software package SPSS 22.0 was used to perform the analyses on a corpus of authentic writing samples from native and nonnative English writers.

Sampling Procedures

The data collection and sampling process took place in September 2017. The writing samples that comprised the corpus in this study were assembled from two major groups of writers. The first group of essays was written by NESs ($n = 100$). These essays were evenly distributed between American and British English speakers. The second major group of essays was from NNEs ($n = 600$), which included six mother tongue groups. These groups were Chinese, German, Japanese, Russian, Spanish, and Turkish. Each language group contributed

100 writing samples to the entire nonnative English writing subcorpus (Table 23). Therefore, the sampling resulted in a total number of 700 essays, with 100 essays from NESs and 600 essays from NNEs, including 100 for each of the six mother tongue groups.

Table 23

Number of Essays by Language Designation

	Number of Essays	Percent in Sample
Nonnative Speakers	600	85.71
Native Speakers	100	14.29
Total	700	100

The research questions of the present study required two types of comparison between different groups of writers to be conducted. First, using NSs as a referential variety, the comparison between native and nonnative English writers provided evidence to demonstrate NNSs' general preparedness for lexical demands of college-level academic writing. Second, the comparison between all seven groups of writers allowed the examination of unique lexical features of each group of writers.

All writing samples were extracted from the ICLE and the LOCNESS. (Detailed descriptions of these two corpora can be found in Chapter Three.) The writing samples in these two corpora were collected from writers with similar ages and grade levels. In terms of the native English writing from the LOCNESS, the writers were American university students and British pupils and university students. To conduct meaningful comparisons, only university level writing samples were selected. Regarding the nonnative English writings from the ICLE, the writers were university students who spoke English as a foreign language. In addition, all nonnative English writers were identified as advanced or high-intermediate English proficiency through consistent testing. The rationale for selecting the students with higher English proficiency is that

advanced ELs are the major population that has the essential and urgent demand of improving academic English writing skills.

The genre of the writing samples extracted for the current study was argumentative writing. The choice of this particular genre of writing is appropriate for the current study because argumentative writing is the common requirement across most of the disciplines in higher education. Hence, to control the scope of the present study while revealing a representative picture of academic writing, argumentative writing is a legitimate option. Accordingly, data collection yielded a suitable corpus of argumentative essays ($N = 700$) for lexical analyses in the present study.

Descriptive Data Results

Descriptive data and measures of central tendency indicated that the mean text length of the essays was 609.03 words ($SD = 224.84$; range, 186-1910; Table 24). For both research questions in the present study, the IV was the language backgrounds of the writers. In terms of the first type of comparison (i.e., Research Question One), two levels of IV were included, namely the native and nonnative English writers. In terms of the second type of comparison (i.e., Research Question Two), seven levels of IV were involved, namely the six groups of NNSs (i.e., Chinese, German, Japanese, Russian, Spanish, and Turkish writers) and the group of NSs. The mean text length of each group of writers is illustrated in Table 24.

Table 24

Text Length

	Language background	<i>N</i>	Range	Minimum	Maximum	<i>M</i>	<i>SD</i>
Nonnative	Chinese	100	767	384	1151	547.21	104.67
	German	100	1055	191	1246	469.62	212.86
	Japanese	100	610	399	1009	569.90	121.46
	Russian	100	1081	186	1267	642.40	261.44
	Spanish	100	929	224	1153	613.15	157.07
	Turkish	100	544	505	1049	744.45	150.70
Nonnative total		600	1081	186	1267	597.79	195.26
Native	English	100	1693	217	1910	676.50	347.67
Total Corpus		700	1724	186	1910	609.03	224.84

Table 24 demonstrates that text length varied largely in the corpus data. Even though the measures of the lexical features have been tested to be independent from text length in previous research (Durán et al., 2004; McCarthy & Jarvis, 2010; McNamara et al., 2005), in order to control for the possible influence from text length, it was set as the covariate in the statistical analyses.

The DVs were divided into three major lexical features, namely lexical diversity, lexical sophistication, and cohesion. The MTLD value was used to represent the level of lexical diversity; coverages of the first 1,000 words, the second 1,000 words, the AWL, the AVL, and off-list words were used to reveal the level of lexical sophistication. Detailed descriptions of these measurements can be found in Chapter Two. In sum, higher MTLD value indicates higher level of lexical diversity; however, higher coverage of the first and second 1,000 words reveals lower level of lexical sophistication. In addition, higher coverage of academic vocabulary (as measured by the AWL and the AVL) and off-list words represents higher level of lexical sophistication.

Three constructs were used to measure cohesion. These three constructs, including referential cohesion, latent semantic analysis (LSA), and connectives, were concluded from the process of Exploratory Factor Analysis (EFA), Chapter Three described the rationale and the process of conducting the EFA as well as the meaning of each construct in depth. In short, higher value of referential cohesion indicates higher level of word repetition and morphological cohesion; while higher LSA value is the reflection of vocabulary association and overlap at the semantic level. For connectives, namely transitional words and phrases, it is critical to employ them in the texts to achieve cohesion; however, overusing connectives may not necessarily lead to higher level and quality of cohesion. Granger and Tyson (1996) found that NNSs tended to overuse connectors when compared to NSs. In addition, Crossley and McNamara's study (2010) revealed that the expert raters in their study positively evaluated the coherence based on the absence of the connectives rather than their presence in the essays. Table 25 elucidates the means, standard deviations, ranges and other descriptive statistics of all the DVs mentioned previously.

Table 25

Descriptive Statistics for Lexical Features

Lexical features		<i>M</i>	<i>SD</i>	Range	Min	Max	Trimmed <i>M</i>	Skewness	Kurtosis	95% CI	
										LL	UL
Lexical diversity (MTLD)		82.33	21.34	148.71	40.81	189.53	81.32	.86	1.68	80.75	83.91
	1st 1000 words	.83	.05	.31	.64	.95	.83	-.47	.22	.83	.84
	2nd 1000 words	.05	.03	.79	.01	.80	.05	14.00	295.71	.05	.06
Lexical sophistication	AWL	.04	.03	.26	.00	.26	.04	1.60	6.94	.04	.04
	AVL	.11	.05	.29	.01	.29	.11	.71	.44	.11	.11
	Off-list words	.07	.06	.99	.00	1.00	.07	9.52	143.40	.07	.08
Cohesion	Referential cohesion	0	.97	5.69	-2.37	3.32	-.01	.31	.41	-.07	.07
	LSA	0	.95	6.24	-2.99	3.25	-.02	.37	.61	-.07	.07
	Connectives	0	1.05	7.23	-2.76	4.48	-.03	.46	.68	-.07	.07

Table 25 shows that some descriptive statistics, such as kurtosis, displayed some abnormal values. Thus, before further analyzing the data with statistical methods, initial data screening was conducted to ensure that the data were appropriate for further inferential statistical analyses. Categorical IV and continuous DVs were both examined.

Initial Data Screening of the Independent Variable

The IV in the present study was the language backgrounds of the writers, which was a categorical variable. In the comparison between native and nonnative writers, nonnative writers were coded as 1 and native writers were coded as 2. Frequency analysis revealed that there was no missing data and all cases were coded as either 1 or 2. In the comparison across all writers from various language backgrounds, Chinese writers were coded as 1, German writers were coded as 2, Japanese writers were coded as 3, Russian writers were coded as 4, Spanish writers were coded as 5, Turkish writers were coded 6, and English writers were coded as 7. Results of the frequency analysis suggested that no missing data were found and each language group contained the same number of cases.

Initial Data Screening of the Dependent Variables

Two MANCOVAs were conducted to examine the two types of comparison. All DVs were continuous data. The initial data screening for the DVs included outlier analysis, skewness, kurtosis, and other normality checks.

First, multivariate outliers were identified with Mahalanobis distance. The multivariate outliers are observations that are inconsistent with the correlational structure of the dataset (Allen, 2017). In addition, since some values of the lexical features had to be manually entered to SPSS, detecting multivariate outliers based on Mahalanobis distance is beneficial for identifying

potential false entry. As a result, eight outliers in the dataset were detected. The descriptive statistics of the eight cases are presented in Table 26.

Table 26

Multivariate Outliers Identified Based on Mahalanobis Distance

ID	Language Group	Text Length	Lexical Diversity	1st 1000	2nd 1000	AWL	AVL	Off-list	Referential Cohesion	LSA	Connectives
39	German	753	189.527	.7813	.1014	.0264	.0712	.0909	-.059	-1.698	.254
59	German	573	65.314	.7535	.0704	.0475	.096	.1285	-.518	2.028	4.481
138	Spanish	762	126.320	.7815	.0609	.0808	.200	.7680	-1.192	.946	.900
160	Turkish	881	60.921	.9057	.0307	.2610	.0736	.0375	-.461	.414	-.022
196	English	730	78.129	.8022	.7970	.0412	.0825	.0769	-1.038	-.438	-1.470
205	English	617	79.480	.7954	.0327	.0720	.1524	.9980	-.537	.827	-.639
431	German	196	102.4700	.7538	.1333	.0051	.0258	.1077	2.182	-2.987	-.748
546	Russian	825	89.088	.7946	.0770	.0257	.0705	.1027	2.352	2.953	-.178

Compared to descriptive statistics presented in Table 24 and Table 25, these eight cases indeed revealed considerable deviation from the means in different DVs. Essay 39, 138, and 431 displayed much higher levels of lexical diversity. Essay 160 demonstrated extremely high coverage of the AWL. For essay 59, its coverage of off-list words was noticeably higher than the mean. Moreover, the values of referential cohesion and LSA for essay 59 and 546 were remarkably higher than the mean as well. Lastly, the coverage of the second 1,000 words in essay 196 and the coverage of off-list words in essay 138 and 205 can be determined as false entry. After removing the eight outliers, the sample size reduced to $N = 692$ (Table 27).

Table 27

Number of Essays after Removing Outliers Based on Mahalanobis Distance

Language background		Frequency	Percent of Sample
Nonnative	Chinese	100	14.5
	German	97	14.0
	Japanese	100	14.5
	Russian	99	14.3
	Spanish	99	14.3
	Turkish	99	14.3
Native	English	98	14.2
Total		692	100.0

Next, the skewness and kurtosis of the DVs under each factor of the IV were evaluated. Significance tests of skewness and kurtosis are not applicable to studies with large sample size because the result is likely to be significant even when the skewness and kurtosis of the data are not too different from normal distribution (Field, 2013). Hence, the present study applied rules of thumb to evaluate the skewness and kurtosis of the data.

For skewness and kurtosis values within the range of ± 2 , normal distribution can be accepted (Gravetter & Wallnau, 2014). The results for each DV are revealed in Table 28 and Table 29.

Table 28

Skewness of Dependent Variables

	NS	NNS	Chinese	German	Japanese	Russian	Spanish	Turkish
Lexical diversity (MTLD)	.700	.755	.157	.832	.434	.413	.668	.341
1 st 1000	.210	-.664	-.394	-.109	-.831	-.759	-.087	-.309
2 nd 1000	.697	.542	.424	.464	.583	.533	.262	.729
AWL	.407	1.080	.607	1.122	1.572	.857	1.010	1.592
AVL	.144	.838	.205	.506	.842	.1179	.298	.702
Off-list	-.070	.789	.164	.474	1.215	.937	.389	.579
Referential cohesion	-.017	.504	.142	.535	.131	.813	.550	.199
LSA	.414	.431	.746	.658	.346	.413	.636	.436
Connectives	.759	.319	.287	-.269	.368	.362	.400	.522

Table 29

Kurtosis of Dependent Variables

	NS	NNS	Chinese	German	Japanese	Russian	Spanish	Turkish
Lexical diversity (MTLD)	.930	1.052	-.318	1.008	-.836	.124	.471	-.680
1 st 1000	.127	.648	-.407	-.363	.557	2.020	-.361	.191
2 nd 1000	-.283	.227	.382	.266	-.343	.231	-.546	.632
AWL	-.014	1.374	-.112	1.458	2.811	.310	2.087	5.223
AVL	-.155	.762	-.745	-.220	.953	1.602	-.023	.601
Off-list	.069	.577	-.567	.021	1.610	2.638	-.027	.123
Referential cohesion	-.627	.569	.362	1.356	-.487	2.058	-.185	-.055
LSA	-.054	.448	-.205	1.113	-.730	.155	1.358	-.237
Connectives	.733	.344	-.358	.134	.304	.286	.110	1.258

It can be seen from the results of skewness and kurtosis that all skewness values and most of the kurtosis values were within or around the range of ± 2 . However, the kurtosis values of Japanese and Turkish writers' AWL coverages were relatively high, meaning the distribution of Japanese and Turkish writers' AWL coverages displayed leptokurtic distribution. French, Macedo, Poulsen, Waterson, and Yu (2008) suggest that *F*-test is robust to non-normality if it is

caused by skewness rather than outliers. In this case, the leptokurtic distribution indicated potential extreme values in the dataset. Due to the relatively minor positive kurtosis of a limited variables, transforming data was avoided to prevent potential harm of generalizing the data. Thus, the extreme values were identified and removed from the corpus. Based on the report of skewness and kurtosis, 11 cases were detected and analyzed (Table 30).

Table 30

Outliers Identified Based on Skewness and Kurtosis

ID	Language group	Text Length	Lexical diversity	1st 1000	2nd 1000	AWL	AVL	Off-list	Referential cohesion	LSA	Connectives
106	Russian	1203	101.062	.7755	.0467	.0609	.1406	.1169	-.884	1.007	-1.059
170	Turkish	514	72.309	.8655	.0253	.0702	.1454	.0390	-.616	-.921	-1.048
295	English	732	70.541	.7717	.0224	.0924	.1562	.1134	.905	.920	-.101
314	Japanese	454	87.674	.7483	.0486	.0795	.1477	.1236	-.935	-.005	-1.120
330	Russian	268	80.787	.7286	.0743	.0706	.1477	.1264	-1.549	.290	-1.774
453	Japanese	553	47.417	.8391	.0235	.1049	.2454	.0325	.875	2.440	1.765
460	Japanese	434	76.166	.6929	.0405	.1048	.1085	.1619	-.052	1.219	-.778
539	Russian	423	126.538	.6905	.0262	.1071	.2743	.1762	1.391	.620	-.978
603	Turkish	890	70.109	.7984	.0315	.1295	.1665	.0405	.423	1.036	.099
615	Turkish	1006	57.421	.8710	.0350	.0700	.0998	.0240	.291	-.048	1.151
622	Turkish	825	72.154	.8262	.0486	.0778	.2305	.0474	.474	.530	.567

As shown in Table 29, Japanese and Turkish writers' coverages of the AWL displayed leptokurtic distribution ($Kurtosis_{\text{Japanese AWL}} = 2.811$; $Kurtosis_{\text{Turkish AWL}} = 5.223$). Therefore, the extreme high values of Japanese and Turkish writers' AWL coverages were identified as essay 170, 314, 453, 460, 603, 615, 622. In addition, in Table 29, Russian writers' coverage of off-list words presented positive kurtosis ($Kurtosis_{\text{Russian off-list}} = 2.638$). Hence, the extreme values of Russian writers' coverage of off-list words were identified as outliers (i.e., essay 106, 330, 539). For essay 295, its coverages of the AWL and the AVL demonstrated apparent deviation from the mean.

As a result, these 11 essays can be considered as extreme values which interfered with the normal distribution of the DVs for each group of writers. To obtain more robust results, these 11 outliers were removed from the corpus, which reduced the final sample size to $N = 681$, total running words changed to 411,845 (Table 31). The final result revealed that all skewness and kurtosis values were within or around the range of ± 2 (Table 32). The resulting sample size of 681 was acceptable for finding a medium effect size ($\alpha = .99$) at the $p < .05$ level (Cohen, 1988).

Table 31

Number of Essays after Removing Outliers based on Skewness and Kurtosis

Language background		Frequency	Percent of Sample
Nonnative	Chinese	100	14.7
	German	97	14.2
	Japanese	97	14.2
	Russian	96	14.1
	Spanish	99	14.5
	Turkish	95	14.0
Native	English	97	14.2
Total		681	100.0
Tokens		411845	100.0

Table 32

Final Descriptive Statistics of Skewness and Kurtosis

	Dependent Variables	NS	NNS	Chinese	German	Japanese	Russian	Spanish	Turkish
Skewness	Lexical diversity (MTLD)	.684	.755	.157	.832	.467	.421	.668	.321
	1st 1000	.191	-.634	-.394	-.109	-.771	-.047	-.087	-.321
	2nd 1000	.691	.535	.424	.464	.556	.563	.262	.703
	AWL	.434	1.053	.607	1.122	1.071	.790	1.010	.769
	AVL	.160	.823	.205	.506	.611	1.057	.298	.492
	Off-list	-.051	.767	.164	.474	1.270	.119	.389	.548
	Referential cohesion	.004	.513	.142	.535	.130	.823	.550	.238
	LSA	.441	.439	.746	.658	.324	.478	.636	.463
	Connectives	.757	.315	.287	-.269	.359	.334	.400	.531
Kurtosis	Lexical diversity (MTLD)	.915	1.065	-.318	1.008	-.778	.197	.471	-.752
	1st 1000	.125	.667	-.407	-.363	.517	.066	-.361	.277
	2nd 1000	-.285	.231	.382	.266	-.392	.295	-.546	.625
	AWL	.044	1.388	-.112	1.458	.685	.038	2.087	1.591
	AVL	-.163	.746	-.745	-.220	.058	1.127	-.023	-.090
	Off-list	.070	.589	-.567	.021	2.076	-.245	-.027	.031
	Referential cohesion	-.619	.589	.362	1.356	-.473	2.390	-.185	-.055
	LSA	.007	.470	-.205	1.113	-.772	.325	1.358	-.202
	Connectives	.697	.378	-.358	.134	.396	.331	.110	1.247

Finally, histograms, Q-Q plots, and previous skewness and kurtosis values were all taken into consideration to check for normality from a more general perspective. The histograms and Q-Q plots of the DVs all revealed approximate normal distributions. Thus, considering all tests and graphs in combination, the DVs in each groups of writers met the requirement of normal distribution.

Assumption Tests and Final Data Screening

After some cases were removed, initial data screening has demonstrated that the DVs reflected approximate normality. To continue with the MANCOVA, a few additional assumptions need to be met to ensure the robustness of the results. In this section, final data screening was conducted respectively to the DVs and each assumption of MANCOVA was tested.

Besides the assumption of normality, assumptions of homogeneity of variance-covariance matrices across cells and linearity need to be checked before conducting MANCOVA. In terms of the homogeneity of variance-covariance matrices, Levene's test was used for each DV to test the equality of variance across the cells. In addition, Box's *M* test was used to test the equality of covariance matrices across the cells.

When the IV was set as two groups of writers, namely the native and nonnative group, for answering the first research question, Levene's test indicated that the assumption of the equality of variances for the coverages of the first 1,000 words, the second 1,000 words, the AWL, the AVL, and off-list words was met with *p* value larger than .05 (Table 33). Consequently, for part of the first research question, which examines the levels of lexical sophistication, the requirements for equal variances between the DVs were fulfilled. Thus, Type I error can be eliminated from the results.

However, the assumption of the equality of variances for measures of lexical diversity and three measures of cohesion was not met with p value smaller than .05 (Table 33). Therefore, for the findings of the differences in lexical diversity and cohesion, it is recommended to interpret cautiously due to the differences in variance between the two groups of writers.

Table 33

Levene's Test with Two Language Groups (Native and Nonnative)

	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i>
Lexical diversity (MTLD)	5.155	1	679	.023
First_1000	3.179	1	679	.075
Second_1000	.968	1	679	.326
AWL	3.601	1	679	.058
AVL	.566	1	679	.452
Off_list	.059	1	679	.809
Referential cohesion	6.465	1	679	.011
LSA	14.426	1	679	.000
Connectives	4.173	1	679	.041

Moreover, when the IV was set as seven language groups of writers for answering the second research question, Levene's test revealed that only the coverage of the second 1,000 words and the DV of connectives demonstrated the equality of variance across cells with p value larger than .05 (Table 34). The unequal variances between most of the DVs might indicate the potential danger of a Type I error. Thus, the interpretation of the findings should be read with caution. To strengthen the reliability of the results, the critical value of α was changed from the conventional .05 to .01.

Table 34

Levene's Test with Seven Language Groups

	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i>
Lexical diversity (MTLD)	3.466	6	674	.002
First_1000	8.705	6	674	.000
Second_1000	1.768	6	674	.103
AWL	9.364	6	674	.000
AVL	12.444	6	674	.000
Off_list	9.991	6	674	.000
Referential cohesion	3.231	6	674	.004
LSA	7.038	6	674	.000
Connectives	1.744	6	674	.108

Regarding Box's *M* test, dividing the IV into either two or seven groups, the *p* values were both smaller than .01, indicating the null hypothesis of equal covariance matrices was rejected. Even though the assumption of homogeneity was violated, research has suggested the robustness of MANCOVA even when homogeneity of variance was violated (Salkind, 2010). Thus, it was still appropriate to continue with further inferential statistical analyses.

Next, the assumption of linearity was tested by conducting scatterplots and correlation matrix to screen the relationships between the DVs. The scatterplots and correlation matrix indicated that most of the DVs demonstrated fair or strong correlations between each other, in particular between the coverages of the first 1,000 words and off-list words as well as between the coverages of the AWL and the AVL (Table 35). The few relatively weak correlations can be found between the cohesion features and other surface level lexical features (i.e., lexical diversity and lexical sophistication), such as the correlation between referential cohesion and the coverage of the AWL as well as between the referential cohesion and the connectives. Since the weak correlations were occasional and the analyses of the lexical features were distinguished, the linearity assumption of the dataset was not considered as being violated.

Table 35

Correlation (Pearson's r) between Dependent Variables

	Lexical diversity	1st 1000	2nd 1000	AWL	AVL	Off_list	Referential cohesion	LSA	Connectives
Lexical diversity	1								
1st 1000	-.252**	1							
2nd 1000	.138**	-.343**	1						
AWL	.089*	-.569**	-.269**	1					
AVL	.009	-.405**	-.208**	.758**	1				
Off_list	.213**	-.847**	.098*	.281**	.175**	1			
Referential cohesion	-.290**	.093*	-.107**	-.014	.015	-.054	1		
LSA	-.424**	-.220**	.026	.291**	.369**	.112**	.056	1	
Connectives	-.037	.261**	.041	-.254**	-.181**	-.233**	.002	-.021	1

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

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

In sum, the homogeneity assumption was partially violated and the linearity assumption was met for the dataset. Nonetheless, MANCOVA has been tested as robust to violation of assumptions. Thus, final data screening indicated that it was appropriate to continue with the MANCOVAs for the two research questions. Considering violations of assumptions existed in the data, significance level was determined as $\alpha = .01$ rather than the conventional .05 in order to further improve the reliability and validity of the results.

Research Questions One

Research Question One was addressed by conducting the first MANCOVA to examine the mean differences between native and nonnative writers' lexical features in academic writing. For lexical diversity, the null hypothesis failed to be rejected, which was reflected as nonsignificant differences regarding lexical diversity between native and nonnative speakers' writing. For lexical sophistication, the null hypothesis was rejected by revealing the significant differences between native and nonnative speakers' writing. In terms of cohesion, the null

hypothesis failed to be rejected, meaning there were no significant differences in cohesion between native and nonnative writers' texts. Detailed statistical analyses are provided in this section.

The group variable, namely IV, was the language designation (NNS = 1, NS = 2); the DVs included three major lexical features. These three lexical features were measured by nine specific DVs, namely lexical diversity, the coverages of the first 1,000 words, the second 1,000 words, the AWL, the AVL, and off-list words, referential cohesion, LSA, and connectives. Results of the multivariate tests demonstrated statistically significant difference in lexical features between native and nonnative writers ($F_{9,670} = 15.325, p < .001, \eta^2 = .171$). Pillai's Trace (.171) was used to interpret the effect size due to the violation of homogeneity of variance (Tabachnick & Fidell, 2007). The effect size can be interpreted as that language designation accounted for 17.1% of the difference between native and nonnative English writers' lexical features in general.

When the nine DVs were considered separately in the MANCOVA, five measures of lexical sophistication revealed significant differences between the two groups of writers (Table 6). Native writers had significantly higher coverages of the AVL, the AWL, and off-list words than did nonnative writers. Meanwhile, nonnative writers' coverages of the high-frequency words, including the first and second 1,000 words, were significantly higher than those of native writers. This indicated that native writers had higher level of lexical sophistication than did nonnative writers. To be specific, tests of between-subjects effects for measures of lexical sophistication revealed that language designation accounted for more than 6% of the variance in the coverage of the first 1,000 words, around 3% in the coverage of the second 1,000 words,

12% in the coverage of the AWL, more than 3% in the coverage of the AVL, and 4.6% in the coverage of off-list words (Table 36).

Table 36

Tests of Between-Subjects Effects for Measures of Lexical Sophistication (MANCOVA 1)

Dependent Variable	Native		Nonnative		<i>df</i>	<i>F</i>	<i>p</i>	η^2
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
1st 1000	.804	.041	.839	.049	(1, 678)	46.119	<.001	.064
2nd 1000	.045	.018	.055	.020	(1, 678)	19.609	<.001	.028
AWL	.062	.026	.038	.022	(1, 678)	95.325	<.001	.123
AVL	.131	.045	.106	.049	(1, 678)	23.213	<.001	.033
Off-list	.088	.033	.068	.033	(1, 678)	32.741	<.001	.046

In terms of the comparisons in lexical diversity, referential cohesion, LSA, and connectives, the MANOVA did not demonstrate statistically significant differences (all $p > .05$, Table 37).

Table 37

Tests of Between-Subjects Effects for Lexical Diversity and Measures of Cohesion

(MANCOVA 1)

Dependent Variable	Native		Nonnative		<i>df</i>	<i>F</i>	<i>p</i>	η^2
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
Lexical diversity (MTLD)	84.439	17.203	81.827	21.514	(1, 678)	2.037	.154	.003
Referential cohesion	.204	1.239	.032	.906	(1, 678)	2.501	.114	.004
LSA	.150	.629	-.041	.977	(1, 678)	2.205	.138	.003
Connectives	-.166	.915	.029	1.058	(1, 678)	3.258	.072	.005

However, because of the unequal variances between the DVs that were mentioned previously (Table 33), the inferential findings should be interpreted cautiously. Hence, mean differences revealed through descriptive statistics (Table 37) were examined to complement the findings. These statistically nonsignificant results demonstrated that native writers' lexical

diversity level was higher than that of nonnative writers on average. In addition, native writers' LSA, namely semantic cohesion, was higher than that of nonnative writers. Nonetheless, native writers' referential cohesion and use of connectives were lower than those of nonnative writers. This means that native writers were able to use varied vocabulary with related meanings to create the deeper level cohesion within the texts, whereas nonnative writers tended to use same words or transitional words and phrases between sentences to form the surface level cohesion within the texts.

In sum, the comparison between native and nonnative English writers demonstrated native writers' higher levels of lexical diversity and lexical sophistication, than those of nonnative writers. In particular, the difference in lexical sophistication was statistically significant. To address the comparison regarding cohesion, the findings failed to reveal the statistically significant difference between native and nonnative writers although descriptive statistics indicated nonsignificant differences (Table 37). The mean differences showed that texts of native English writers had higher level of semantic cohesion than those of nonnative writers. However, NNEs' writing showed higher level of referential cohesion (i.e., word repetition) and more incidences of connectives (i.e., transaction words). This result is in line with Flowerdew (1998) which also notes NNSs' overuse of connectors at the local rather than the global level. Crossley and McNamara (2011) also suggest a negative correlation between the use of cohesive devices (i.e., connectives) and the writers' proficiency level.

Research Questions Two

To answer the second research question, the second MANCOVA was conducted to reveal further details in terms of the differences across all groups of writers from various language backgrounds. All three hypotheses were rejected as there were significant differences between

writers from various language backgrounds in all three lexical features in academic writing, namely lexical diversity, lexical sophistication, and cohesion.

As with the first MANCOVA, the IV was still the language designation. However, the IV in the second MANCOVA contained seven levels, namely the seven different mother tongue backgrounds, including Chinese, German, Japanese, Russian, Spanish, Turkish, and English. The DVs were the same as in the first MANCOVA, which included nine specific measurements for three major lexical features.

Also, Pillai's Trace (1.010) was employed to interpret the results because of the violation of homogeneity of variance (Tabachnick & Fidell, 2007). Results of the multivariate tests indicated a significant difference in the combined measures of lexical features across different groups ($F_{54, 4020} = 15.061, p < .001, \eta^2 = .168$). Language backgrounds explained 16.8% of the variance in the differences of the lexical features in general.

Tests of between-subjects effects revealed more detailed comparison in terms of each measure of the lexical features based on language backgrounds. Different from the first MANCOVA, the results revealed significant differences across different groups of writers with regard to all nine measures of the lexical features (Table 38).

In sum, the second research question can be answered by rejecting the null hypotheses and revealing the significant differences between all seven groups of writers in terms of three major lexical features.

Table 38

Tests of Between-Subjects Effects for All Lexical Features (MANCOVA 2)

Dependent Variables	<i>df</i>	<i>F</i>	<i>p</i>	η^2
LDMTLD	(6, 673)	41.031	<.001	.268
AVL	(6, 673)	38.558	<.001	.256
First_1000	(6, 673)	33.553	<.001	.230
Second_1000	(6, 673)	11.237	<.001	.091
AWL	(6, 673)	41.511	<.001	.270
Off_list	(6, 673)	26.020	<.001	.188
Referential cohesion	(6, 673)	9.402	<.001	.077
LSA	(6, 673)	46.001	<.001	.291
Connectives	(6, 673)	4.125	<.001	.035

To deepen the analyses, in the next section, a series of post hoc analyses were conducted to reveal the particular groups of writers who had significant differences between each other in various lexical features. With the specific comparisons, practical significance of the differences presented between various groups of writers were addressed.

Post Hoc Analyses

The analyses presented in answer to Research Question Two suggested statistically significant differences were evident in all three lexical features between writers from seven different language backgrounds. This section provides a detailed *post hoc* analysis of those differences.

Results from the measures of lexical diversity and lexical sophistication revealed that NESs had a relatively high level of lexical diversity and were able to employ more academic vocabulary and low-frequency words, which indicated that their lexical sophistication level was relatively high as well. Among nonnative writers, Chinese writers were able to use more diverse vocabulary at the same time employ more low-frequency words and academic vocabulary. In contrast, Turkish and Japanese writers overused high-frequency words and underused academic

and off-list words. For German writers, even though their lexical diversity level was the highest, their coverage of the academic words was the lowest. Russian and Spanish writers' performances were in the average places regarding all measures.

In terms of cohesion, NESs used relatively fewer repetitive words and connectives; however, their writing presented relatively high level of cohesion at the semantic level. Among nonnative writers, Japanese writers performed well in all three types of cohesion; in contrast, Russian writers had low levels of all cohesion constructs, namely referential cohesion, LSA, and connectives. Chinese writers had high level of content words overlapping (i.e., word repetition) and were able to build cohesion at the semantic level; nonetheless, they employed fewer connectives. For Turkish and German writers, even though they were able to use more connectives, their vocabulary and semantic level cohesion were relatively low compared to other groups of writers. Spanish writers were capable of building cohesion in vocabulary overlapping; however, they failed to keep the high cohesion level in the semantic aspect. Next, detailed statistical analyses of the pairwise comparisons are provided to support the summarization.

Lexical Diversity

For lexical diversity, Table 39 elucidates the means of each language group with the order from low to high. The pairwise comparison (Table 40) indicated that Japanese and Turkish writers' lexical diversity were significantly lower than all other groups of writers (all $p < .01$). German and Russian speakers' writing had the highest lexical diversity among all groups of writers, and the differences between German and Russian writers and other groups of writers were statistically significant (all $p < .01$). This result indicates that compared to other NNSs, Japanese and Turkish writers might have difficulty in diversifying their vocabulary in academic English writing, whereas German and Russian speakers' face less challenge in this aspect.

Table 39

Descriptive Statistics of Lexical Diversity

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
Japanese	97	66.263	15.565
Turkish	95	71.092	14.413
Chinese	100	78.399	16.319
Spanish	99	82.453	18.702
English	97	84.439	17.203
Russian	96	91.565	18.689
German	97	101.161	23.320
Total	681	82.199	20.963

Table 40

Pairwise Comparisons (Mean Difference) of Lexical Diversity

	CH	GE	JA	RU	SP	TU
GE	22.774*					
JA	-12.141*	-34.915*				
RU	13.151*	-9.623*	25.291*			
SP	4.044	-18.730*	16.184*	-9.107*		
TU	-7.339*	-30.113*	4.802	-20.489*	-11.382*	
EN	6.019	-16.755*	18.159*	-7.132*	1.975	13.357*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Coverage of the First 1,000 Words

In terms of the coverage of the first 1,000 words (Table 41 & Table 42), Chinese and native English writers had the lowest coverages and they were both significantly lower than other groups of writers (all $p < .001$). Japanese and Turkish writers had the highest coverages, in particular, Turkish writers' coverage of the first 1,000 words was significantly higher than other five groups of writers (all $p < .01$). This result demonstrates that Turkish and Japanese writers might need extra assistance in improving their vocabulary size beyond the high-frequency level words.

Table 41

Descriptive Statistics of the Coverage of the First 1,000 Words

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
Chinese	100	.798	.058
English	97	.804	.041
German	97	.829	.041
Russian	96	.842	.033
Spanish	99	.840	.041
Japanese	97	.856	.052
Turkish	95	.869	.034
Total	681	.834	.050

Table 42

Pairwise Comparisons (Mean Difference) of the Coverage of the First 1,000 Words

	CH	GE	JA	RU	SP	TU
GE	.031*					
JA	.058*	.027*				
RU	.044*	.013	-.014			
SP	.042*	.011	-.016	-.002		
TU	.071*	.040*	.013	.027*	.029*	
EN	.006	-.025*	-.051*	-.038*	-.035*	-.065*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Coverage of the Second 1,000 Words

Regarding the coverage of the second 1,000 words (Table 43 & Table 44), English and Spanish writers had the lowest coverages in their writing. English writers' coverage was significantly lower than all other groups of writers (all $p < .01$), except for Spanish writers ($p = .598$). For Spanish writers, their coverage of the second 1,000 words was significantly lower than other three groups of writers (all $p < .01$). Lastly, Chinese and German writers had the highest coverages of the second 1,000 words; in particular, German writers' coverage was significantly higher than all other groups of writers (all $p < .01$), except for Chinese writers (p

= .245). Chinese writers' coverage was significantly higher than the other four groups of writers (all $p < .01$) but not significantly different from Russian writers' ($p = .077$).

Table 43

Descriptive Statistics of the Coverage of the Second 1,000 Words

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
English	97	.045	.018
Spanish	99	.047	.017
Turkish	95	.052	.018
Japanese	97	.053	.022
Russian	96	.055	.017
Chinese	100	.061	.019
German	97	.064	.022
Total	681	.054	.020

Table 44

Pairwise Comparisons (Mean Difference) of the Coverage of the Second 1,000 Words

	CH	GE	JA	RU	SP	TU
GE	.003					
JA	-.008*	-.011*				
RU	-.005	-.008*	.003			
SP	-.014*	-.017*	-.006	-.009*		
TU	-.008*	-.011*	-.001	-.003	.006	
EN	-.015*	-.019*	-.008*	-.010*	-.001	-.008*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Coverage of the High-Frequency Words

Taken together the coverages of the high-frequency words, including coverages of the first and second 1,000 words (Table 45), native English and Chinese writers' coverages were the lowest, both were around 85%. Japanese and Turkish writers' coverages of the high-frequency words were the highest and both reached more than 90%. Spanish, German, and Russian writers'

coverages were all around 89%. Again, the combined result sheds light on the additional demands that Turkish and Japanese writers might need in terms of improving their advanced vocabulary knowledge.

Table 45

Descriptive Statistics of the Coverage of the First 2,000 Words

Language background	<i>N</i>	First 1,000 <i>M</i>	Second 1,000 <i>M</i>	First 2,000 <i>M</i>
English	97	.804	.045	.849
Chinese	100	.798	.061	.859
Spanish	99	.840	.047	.887
German	97	.829	.064	.893
Russian	96	.842	.055	.897
Japanese	97	.856	.053	.909
Turkish	95	.869	.052	.921
Total	681	.834	.054	.888

Coverage of the AWL

In addition to the coverages of the high-frequency words, measures of lexical sophistication also focus on the coverages of the low-frequency words, including academic vocabulary and off-list words. Updated version of the LFP uses the coverage of the AWL to reveal writers' knowledge of academic vocabulary in writing. The second MANCOVA demonstrated that NESs were able to employ the highest coverage of words from the AWL and it was significantly higher than all other nonnative writers (all $p < .01$; Table 46 & Table 47). Chinese writers' coverage of the AWL was the second highest and it was also significantly higher than the other five groups of writers (all $p < .01$). The lowest coverages were from German writers. Their coverage of the AWL was significantly lower than that of all other groups and writers (all $p < .01$), except for Japanese writers ($p = .146$). Japanese writers' coverage of the AWL was the second lowest and was significantly lower than that of English, Chinese, Spanish, and Russian writers (all $p < .01$). In sum, although German writers showed performed averagely

in terms of the coverage of high-frequency words, their low coverage of the AWL revealed their needs in mastering more academic vocabulary. For Japanese writers, besides additional assistance in controlling their use of high-frequency words, further instruction in employing more academic vocabulary is also needed.

Table 46

Descriptive Statistics of the Coverage of the AWL

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
German	97	.024	.017
Japanese	97	.028	.016
Turkish	95	.034	.016
Russian	96	.041	.020
Spanish	99	.044	.020
Chinese	100	.054	.028
English	97	.062	.026
Total	681	.041	.024

Table 47

Pairwise Comparisons (Mean Difference) of the Coverage of the AWL

	CH	GE	JA	RU	SP	TU
GE	-.030*					
JA	-.026*	.004				
RU	-.013*	.017*	.013*			
SP	-.010*	.020*	.016*	.003		
TU	-.019*	.011*	.006	-.007	-.010*	
EN	.009*	.039*	.034*	.021*	.018*	.028*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Coverage of the AVL

Considering the existing criticisms toward the AWL, the coverage of the AVL was also used to help reveal the writers' knowledge of academic vocabulary in writing. Since the AVL employs lemmas rather than word families to compile the list (Gardner & Davies, 2013), not

surprisingly, the mean coverages of the AVL were higher than what the AVL showed ($M_{AWL} = .041$, $SD_{AWL} = .024$; $M_{AVL} = .110$, $SD_{AVL} = .049$). Nonetheless, the rank revealed from the coverages of the AVL was almost the same as the coverages of the AVL, except that the Spanish writers had slightly lower coverage than did the Russian writers and Chinese writers' coverage was higher than that of the native English writers ($p = .017$). Moreover, the significance of the differences between the seven groups of writers shown by the AVL also differed from the AVL coverage in some cases. German writers' coverage of the AVL was significantly lower than all other groups of writers (all $p < .01$). Japanese writers' coverage of the AVL was the second lowest and significantly lower than the other groups of writers, except for the group of Turkish writers ($p = .474$). Native English and Chinese writers' coverages were the highest and both were significantly higher than other five groups of writers (all $p < .01$). Table 48 and Table 49 summarizes the statistics of the coverage of the AVL.

Table 48

Descriptive Statistics of the Coverage of the AVL

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
German	97	.066	.033
Japanese	97	.096	.039
Turkish	95	.098	.035
Spanish	99	.112	.038
Russian	96	.115	.043
English	97	.131	.045
Chinese	100	.148	.061
Total	681	.110	.049

Table 49

Pairwise Comparisons (Mean Difference) of the Coverage of the AVL

	CH	GE	JA	RU	SP	TU
GE	-.084*					
JA	-.052*	.032*				
RU	-.032*	.052*	.020*			
SP	-.036*	.048*	.016*	-.004		
TU	-.047*	.037*	.005	-.016	-.012	
EN	-.015	.069*	.037*	.017*	.021*	.032*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Coverage of Off-list Words

Finally, for the coverage of off-list words (Table 50 & Table 51), German, Chinese and native English writers had the highest coverages among all groups of writers. Turkish writers' coverage was the lowest and it was significantly lower than all other groups of writers (all $p < .01$). Thus, for Turkish writers, besides controlling their use of high-frequency words, additional help might be needed in improving their advanced or discipline-specific vocabulary.

Table 50

Descriptive Statistics of the Coverage of Off-list Words

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
Turkish	95	.045	.019
Russian	96	.062	.022
Japanese	97	.063	.037
Spanish	99	.070	.028
German	97	.081	.030
Chinese	100	.088	.039
English	97	.088	.033
Total	681	.071	.034

Table 51

Pairwise Comparisons (Mean Difference) of the Coverage of Off-list Words

	CH	GE	JA	RU	SP	TU
GE	-.006					
JA	-.025*	-.019*				
RU	-.027*	-.021*	-.002			
SP	-.018*	-.012*	.006	.008		
TU	-.044*	-.038*	-.020*	-.018*	-.026*	
EN	.000	.006	.024*	.026*	.018*	.044*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Referential Cohesion

Next, results of the three measures of cohesion are addressed, including the referential cohesion, LSA, and connectives. In terms of the referential cohesion (Table 52 and Table 53), which refers to exact or related content words overlapping (i.e., word repetition), the result indicated that Spanish writers had the highest level of referential cohesion, the differences between Spanish writers and the other groups of writers were all statistically significant (all $p < .01$). Russian writers had the lowest level of referential cohesion and it was significantly lower than the highest three groups of writers, namely Chinese, Japanese and Spanish writers (all $p < .01$). This result suggested that on average, Russian writers tended not to use too many repetitive or similar words in their writing. On the other hand, in Chinese, Japanese, and Spanish writers' essays, overlaps of content words or pronouns were more frequent.

Table 52

Descriptive Statistics of Referential Cohesion

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
Russian	96	-.367	.829
English	97	-.204	1.239
Turkish	95	-.136	.759
German	97	-.072	.871
Chinese	100	.093	.843
Japanese	97	.117	.841
Spanish	99	.537	1.022
Total	681	-.002	.963

Table 53

Pairwise Comparisons (Mean Difference) of Referential Cohesion

	CH	GE	JA	RU	SP	TU
GE	-.239					
JA	.047	.286				
RU	-.373*	-.134	-.420*			
SP	.504*	.743*	.457*	.877*		
TU	-.049	.190	-.096	.324	-.553*	
EN	-.176	.062	-.224	.196	-.681*	-.128

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

LSA

For the LSA (Table 54 & Table 55), which refers to semantic cohesion between sentences and paragraphs, German writers' level of LSA was the lowest and it was significantly lower than all other groups of writers (all $p < .01$). Spanish and Russian writers' LSA level were also relatively low but they did not demonstrate significant difference between each other ($p = .988$). Both Spanish and Russian writers had significantly lower LSA scores when compared to the other four groups of writers. Japanese and Chinese writers' LSA levels were the highest and they were both significantly higher than the four lowest groups of writers, namely German,

Spanish, Russian, and Turkish writers (all $p < .01$). What is more, Chinese writers' LSA level was significantly higher than that of native English writers ($p < .01$). With the difference between the lowest and the highest level of LSA among the groups, the relatively large gap in terms of the LSA score was noteworthy.

Different from the result of referential cohesion, Spanish writers' LSA was relatively low among the NNSs. This indicated that Spanish writers were aware of making the connection between sentences by repeating the same words; however, it might be more difficult for them to employ varied vocabulary that have related meanings. This issue was also a challenge for German speakers as well.

Table 54

Descriptive Statistics of the LSA

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
German	97	-1.076	.772
Spanish	99	-.185	.658
Russian	96	-.181	.698
Turkish	95	.149	.774
English	97	.150	.629
Japanese	97	.401	.870
Chinese	100	.634	1.033
Total	681	-.013	.937

Table 55

Pairwise Comparisons (Mean Difference) of the LSA

	CH	GE	JA	RU	SP	TU
GE	-1.695*					
JA	-.237	1.458*				
RU	-.832*	.863*	-.595*			
SP	-.830*	.865*	-.593*	.002		
TU	-.521*	1.174*	-.284*	.311*	.309*	
EN	-.508*	1.187*	-.271	.324*	.322*	.013

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

Connectives

Lastly, results of the incidence of connectives (Table 56 & Table 57) showed that connectives occurred the least frequently in Russian speakers' writing; English and Chinese writers' use of connectives followed that of Russian writers. German, Japanese, and Turkish writers were the groups used relatively more times of connectives. Turkish writers had the significantly higher incidences of connectives than all other groups of writers (all $p < .01$), except for Japanese writers.

Combining the results of Russian writers in referential cohesion and LSA, it is not difficult to observe that Russian writers' building of cohesion in all three aspects was not ideal, which was reflected by the lowest coverage of word overlaps (i.e., referential cohesion), relatively low level of semantic connections (i.e., LSA), and lowest use of transitions words and phrases (i.e., connectives). In contrast, Japanese writers' cohesion level in all three aspects was among the highest in nonnative English writers.

Table 56

Descriptive Statistics of Connectives

Language background	<i>N</i>	<i>M</i>	<i>SD</i>
Russian	96	-.238	.924
English	97	-.166	.915
Chinese	100	-.147	1.112
Spanish	99	-.038	.950
German	97	.080	1.001
Japanese	97	.156	1.174
Turkish	95	.370	1.075
Total	681	.001	1.040

Table 57

Pairwise Comparisons (Mean Difference) of Connectives

	CH	GE	JA	RU	SP	TU
GE	.231					
JA	.301	.069				
RU	-.097	-.328	-.397*			
SP	.105	-.127	-.196	.201		
TU	.506*	.274	.205	.602*	.401*	
EN	-.027	-.259	-.328	.069	-.132	-.533*

Based on estimated marginal means

*. The mean difference is significant at the .01 level

CH = Chinese; GE = German; JA = Japanese; RU = Russian; SP = Spanish; TU = Turkish; EN = English

The pair-wise comparison of each lexical feature between various groups of writers demonstrated intricate characteristics of the writers. To present the comparison and different groups of writers' lexical performance in an overall manner, three bar graphs were created to illustrate the differences in lexical diversity, lexical sophistication, and cohesion between seven language groups overall. Figure 6 reveals that Chinese, Japanese, and Turkish writers' lexical diversity had the relatively lower values among the seven groups of writers.

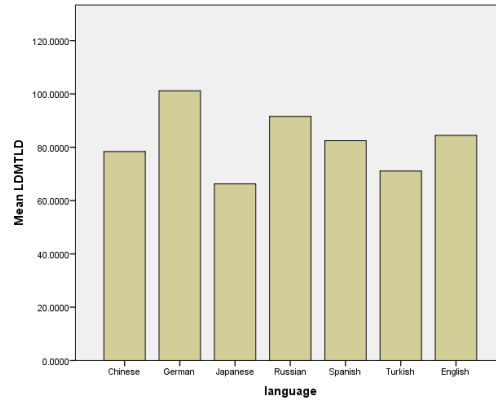


Figure 6: Mean Differences in Lexical Diversity across Writers from Various Language Groups

Figure 7 shows that coverages of the first 2,000 high-frequency words did not present remarkable differences across all groups of writers. However, coverages of academic words, including both the AWL and the AVL, revealed considerable variance across different groups of writers. German, Japanese, and Turkish writers had the lowest coverages of academic words while Chinese writers demonstrated approximately similar coverage of academic words as did native English writers. Regarding the coverages of off-list words, Turkish writers presented apparent lower coverage compared to the other groups of writers.

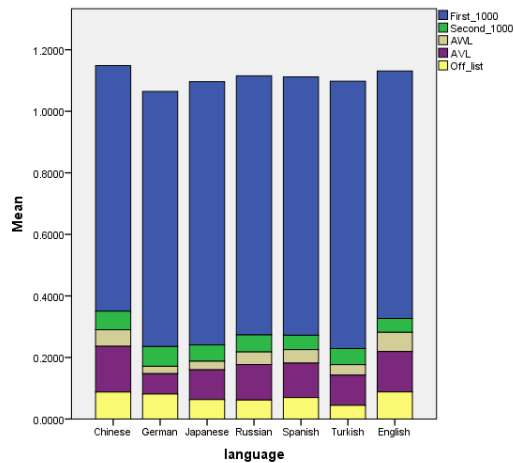


Figure 7: Mean Differences in Lexical Sophistication across Writers from Various Language Backgrounds

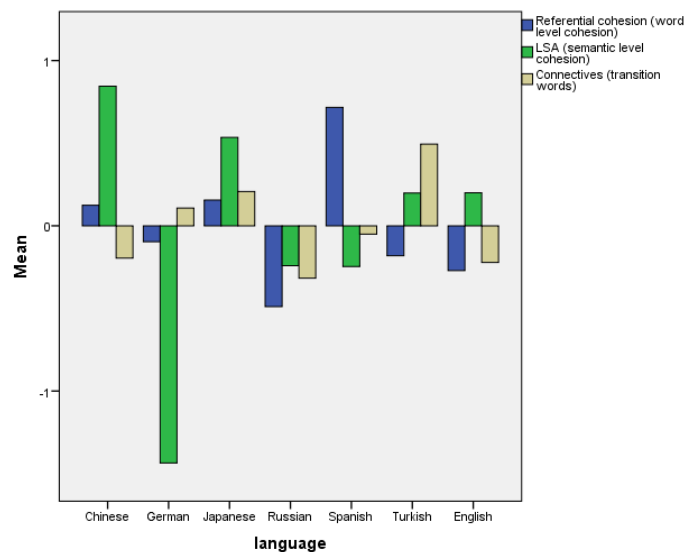


Figure 8: Mean Differences in Cohesion across Writers from Various Language Backgrounds

Finally, it is evident in Figure 8 that German, Japanese, and Turkish writers used more connectives in their writing than other groups of writers. In particular, for German writers, overusing connectives did not improve their levels of referential cohesion and LSA. Russian writers' levels of all three measures of cohesion were relatively low compared to other groups of writers. Different from all nonnative writers, native English writers presented low level of

referential cohesion and scarce use of connectives; however, their level of LSA was relatively high. This indicated that native English writers were able to use approaches other than connectives and repetitive words to create cohesion at a semantic level.

Conclusion

In sum, Chapter Four presented the results of statistical analyses that examined the differences 1) between native and nonnative English writers and 2) across all writers from various language backgrounds in terms of lexical diversity, lexical sophistication, and cohesion in academic writing. Two MANCOVAs were conducted with language designation as the IV, different measures of lexical features as the DVs, and text length as the covariate. The results of the MANCOVAs revealed that NSs were able to employ more diverse vocabulary and had a higher lexical sophistication level than NNSs. In addition, native English writers were able to establish more cohesion semantically (i.e., LSA) than were nonnative English writers. For nonnative English writers, even with higher levels of cohesion in repetitive words (i.e., referential cohesion) and more uses of connectives, they failed to create higher levels of cohesion at the semantic level when compared to native English writers.

Regarding the comparison across all groups of writers, the results revealed that writers from various language backgrounds did demonstrate significant differences between each other. Some groups of writers performed significantly better in lexical diversity and lexical sophistication than others, such as demonstrating more coverages of low-frequency and academic vocabulary; some groups of writers performed significantly better in cohesion than others, such as revealing higher levels of referential cohesion and LSA. The differences presented here indicate the varied characteristics and needs of different groups of NNSs as well as the necessity of target instruction in vocabulary and academic writing.

Chapter Five further interprets the results from the statistical analyses. The limitations of the current study are addressed as well. Combining the interpretation from the statistical results, Chapter Five presents implications for providing targeted vocabulary instruction in academic writing for nonnative English writers from different language backgrounds. Recommendations for further research are provided as the conclusion of the chapter.

CHAPTER FIVE: DISCUSSION

This chapter summarizes the present study which examined the differences in three major lexical features in academic writing 1) between native and nonnative English writers and 2) across all writers from various language backgrounds. The purpose, the major findings, and the limitations of the study are discussed. This chapter concludes with pedagogical implications for NNEs from different language backgrounds and the recommendations for future research.

Purpose of the Study

For any student to succeed in university classes, being able to present one's learning through writing assignments is essential. Across various disciplines, many university-level courses are designed to evaluate students' performance with heavy focus on their writing in the discipline, namely their academic writing performance. Students are required to demonstrate their mastery of the course knowledge by providing clear, convincing, and well-organized academic writing.

Nevertheless, writing has been considered as one of the most challenging aspects for university students to master (Pirttimaa, Takala, & Ladonlahti, 2015). Many university students identify themselves as "bad writers" and often struggle to become more skilled users of academic discourses that are required in college-level classes (Fernsten & Reda, 2011). For NNEs, the struggle in improving academic writing performance is even more (Leki, 2017). NNEs are constantly compared to their native speaking peers and are usually expected to perform at a similar level. Unfortunately, not being able to achieve better performance in writing potentially harms their academic success, resulting in other problems. Thus, assisting NNEs in enhancing their academic writing performance is at the very heart of the current study.

Research in L2 writing has burgeoned in the last few decades (Leki, Cumming, & Silva, 2008). A large number of empirical studies have researched the essential and diverse aspects in teaching and learning L2 writing, such as L2 writing feedback, L2 writers' characteristics, L1 influence on L2 writing, and so forth. Although diverse opinions have been raised in terms of the strategies to improve NNSs' writing performance, almost all researchers and educators agree on the importance of vocabulary knowledge to successful writing, especially writing for academic purposes. Gonzalez (2013) suggests the substantial impact of lexical diversity on writing scores. Omidian, Beliaeva, Todd, and Siyanova-Chanturia (2017) address the importance of employing academic vocabulary to achieve better performance in academic writing. Researchers have also demonstrated the essential role of lexical bundles in academic writing and compared the differences between native and nonnative uses of lexical bundles in writing (Ädel & Erman, 2012).

Hence, it is clear that better knowledge of vocabulary in terms of size, depth, and correct usage helps writers establish convincing examples, clear organization, and strong arguments in their writing, all of which are basic to good academic writing. At the same time, research has shown the struggles that NNSs face regarding improving their vocabulary ability in academic writing (Coxhead, 2012). Different studies have focused on various aspects of vocabulary instruction and learning for NNSs to improve their writing performance, such as lexical errors, lexical diversity, lexical bundles, and so forth. Nevertheless, very few empirical studies have considered the issue of vocabulary in writing from a holistic point of view. In addition, little research has been conducted based on learners' varying needs in the field of L2 writing. Hence, the current study has been conducted to demonstrate a holistic picture of NNSs' vocabulary

performance in academic writing, which provides empirical evidence of learners' unique needs for vocabulary instruction in the field of academic English writing.

Although the ultimate objective for university students in writing is to achieve a very high level of proficiency, NNSs feel the pressure of constantly being compared to their native speaking peers. Consequently, having a clear knowledge of the potential differences between NNSs and their native speaking peers provides teachers and learner themselves with a baseline as well as a realistic goal in improving NNSs' academic writing performance. Hence, the present study first examined the differences in lexical quality between NSs' and NNSs' academic writing. Because of the diversity among NNSs in almost any type of classroom, the current study also considered the characteristics of each language group in academic English writing.

Therefore, the second comparison that the present study conducted was to examine the different lexical performance in academic writing across all writers from various language backgrounds. Furthermore, the three lexical features, namely lexical diversity, lexical sophistication, and cohesion, that the present study examined provided a comprehensive picture of the writers' lexical performance and enriched the existing literature in the field of vocabulary and L2 writing.

As a result, the present study aimed to analyze and compare lexical characteristics of both native and nonnative academic English writing. The corpus-based approach employed in the present study revealed what native and different nonnative writers were able and unable to produce in academic writing in a natural setting. Compared to interviews, self-reports, surveys, and other evaluation approaches, a systematically compiled corpus offers the benefit of having a large number of authentic texts that can represent the writers' characteristics more objectively and in a more generalizable way.

In short, the purposes of the current study were to examine native and nonnative English writers' lexical features in academic writing and to demonstrate 1) the potential differences between NNSs and NSs and 2) the latent differences across writers from various language backgrounds. Based upon corpus data, the varied characteristics of NNSs' lexical features in academic writing provide empirical evidence on various learner needs, which indicates the necessity and directions of targeted vocabulary instruction in L2 academic English writing.

Summary of the Findings

A sample corpus of 700 authentic argumentative essays written by native and nonnative English speakers were extracted from the International Corpus of Learner English (ICLE) and the Louvain Corpus of Native English Essays (LOCNESS) during the Fall semester of 2017. The NNS subcorpus involved nonnative argumentative essays from six different mother tongue groups, including Chinese, German, Japanese, Russian, Spanish, and Turkish speaking ELs. Each language group contributed 100 texts to the subcorpus ($n = 600$). All NNSs were tested with high-intermediate to advanced English proficiency; in addition, the NNSs were at similar age and grade levels in universities where English was taught as a foreign language (Granger et al., 2009). The NS subcorpus involved argumentative essays written by NESs from British and American universities. One hundred NS texts were randomly selected from the LOCNESS ($n = 100$).

Before conducting the comparisons, all texts were entered into computational measuring tools to receive the raw descriptive results of all lexical features, including lexical diversity, lexical sophistication, and cohesion.

Lexical diversity and constructs of cohesion were measured through Coh-Metrix 3.0, which is freely accessible on the Internet. The Measure of Textual Lexical Diversity (MTLD)

value was used to measure the level of lexical diversity. Higher MTLD value refers to higher level of lexical diversity.

The lexical sophistication feature was measured by the Lexical Frequency Profiles (LFP), which evaluates the coverages of the first most frequent 1,000 words in English, the second most frequent 1,000 words, words from the Academic Word List (AWL), and off-list words in the texts. Higher coverages of the AWL and off-list words refer to the ability of employing more low-frequency words, which reflects as higher levels of lexical sophistication. The LFP is embedded in the website of Compleat Lexical Tutor, which is also freely accessible online. Additionally, the coverage of the Academic Vocabulary List (AVL) was measured to complement the examination of academic word usage. The AVL is freely accessible online at www.wordandphrase.info/academic.

Nineteen indices related to cohesion were selected to measure the cohesion levels of the texts. To better conduct the statistical analyses, an Exploratory Factor Analysis (EFA) was carried out to reduce the 19 indices to three representative constructs, namely referential cohesion, Latent Semantic Analysis (LSA), and connectives. These three new constructs can demonstrate the texts' cohesion levels to a large extent (more than 64%) at the same time reduce the intricacy of using all 19 indices for measuring. The referential cohesion describes the content words overlapping (i.e., word repetition) and morphological similarity between sentences and paragraphs in the texts. The LSA depicts the semantic-level connection and cohesion in the texts, which indicates deeper level of cohesion. The connectives are the incidence of employing various types of connectives, namely transition words and phrases.

After obtaining raw results from all lexical measures, two MANCOVAs were conducted to answer the following two research questions. Text length was included as the covariate to

partial out the influence from text length and strengthen the results. The two research questions are provided in the following section and the results of the two MANCOVAs are summarized respectively.

1. Are there significant differences in lexical features between native and nonnative academic English writing, as measured by lexical diversity, lexical sophistication, and cohesion?
2. Are there significant differences in lexical features, as measured by lexical diversity, lexical sophistication, and cohesion, in academic English writing across all writers from various mother tongue backgrounds?

Research Questions One

The first research question targeted at the differences in lexical features between native and nonnative English writers' academic writing. It was hypothesized that NSs' writing would reveal significantly higher levels of lexical performance than that of NNSs in the three lexical features. The first MANCOVA was conducted and demonstrated that NS writing samples indeed exhibited a significant higher level of lexical sophistication than did NNSs' texts, namely all five measures of lexical sophistication have shown the significant differences between NSs and NNSs (all $p < .001$). NSs were able to employ significantly higher coverages of academic vocabulary (as measured by the AWL and the AVL) and low-frequency words (as measured by off-list words) than were NNSs. Unsurprisingly, NNSs' coverages of the high-frequency words, both the first 1,000 and the second 1,000 words, were significantly higher than those of NSs.

On the other hand, in terms of lexical diversity and three measures of cohesion (i.e., referential cohesion, LSA, and connectives), the results did not reveal significant differences between NSs and NNSs. For lexical diversity, NSs indeed performed at a higher level than did

NNSs, however, the difference was not statistically significant. For referential cohesion and connectives, NSs performed at a lower level than did NNSs, whereas NSs' LSA level was higher than that of NNSs. Even though the differences in measures of cohesion failed to reach statistical significance, the findings suggested that NNSs were better at building simple and superficial cohesion, such as word repetition and employing transitional words; nonetheless, NSs appeared to be better at establishing connections and cohesion at a deeper semantic and global level by using limited repetitive vocabulary or without simply inserting transitional words in the texts.

In sum, the findings answered the first research question as there were significant differences between native and nonnative English writers' lexical sophistication level in their academic writing; however, regarding lexical diversity, the two groups of writers were not distinguished with statistically significant differences even though NSs demonstrated a higher level on average than did NNSs. Although the results only revealed statistically significant differences between NSs' and NNSs' performance in lexical sophistication, when compared to NSs, NNSs performed at a lower level in terms of their use of diverse, academic, and low-frequency vocabulary. This result is in accordance with Douglas (2010), Gonzalez (2013), Kwon (2009), Omidian et al. (2017), and Paquot (2010), which all shed light on the substantial differences between native and nonnative writing regarding lexical diversity and lexical sophistication.

For the comparison in cohesion, the null hypothesis was failed to be rejected, meaning there were no significant differences in cohesion between native and nonnative writers' texts. However, the mean differences suggested that NNSs had a higher level of referential cohesion and employed more connectives, whereas NSs had a higher level of LSA. This result is in line with Granger and Tyson's (1996) study, which also noted that NNSs tended to overuse

connectors. Both studies revealed the lack of appropriate semantic and syntactic use in the learner texts. A more recent study by Ma and Wang (2016) also demonstrated the differences between native and nonnative speaking students' use of connectors in writing. In addition, Crossley and McNamara's (2009) study focusing on cognitive level lexical features in native and nonnative writing found a similar pattern in LSA as in the current study, namely native writers' texts were deemed more cohesive through the use of previously given information whereas more new information was apt to be embedded in NNSs' writing. This tendency played a negative role in establishing deeper-level cohesion in nonnative writers' texts.

Research Questions Two

The next research question examined the differences in lexical features of academic writing across all writers from various language backgrounds. It was hypothesized that there would be significant differences between various groups of writers in all three lexical features. The second MANCOVA was then conducted and the results indicated that significant differences indeed existed between at least two groups of writers' lexical features ($F_{54, 4086} = 14.738, p < .001, \eta^2 = .163$). Tests of between-subjects effects also revealed that all three measures of lexical features (i.e., lexical diversity, lexical sophistication, and cohesion) contributed to the significance. Thus, the three null hypotheses of the second research question were rejected.

The results of the pair-wise comparison suggested that Chinese writers' texts ($n = 100$) presented the most similarity to the NSs' writing ($n = 97$), which demonstrated relatively high level of lexical diversity and high coverages of academic and low-frequency vocabulary; meanwhile, compared to other groups of writers, Chinese writers were able to build relatively more word repetition (i.e., referential cohesion) and semantic coreferentiality (i.e., LSA). In contrast, other nonnative writers' essays all demonstrated different strengths and weaknesses in

lexical quality. For instance, Turkish and Japanese writers tended to overuse high-frequency words while underuse academic and off-list words. German speakers' writing had extremely low coverages of academic words when compared to other nonnative writers; additionally, Russian writers presented low cohesion in their writing. Based upon different groups of writers' lexical features, detailed analyses and recommendations are provided in the following section of pedagogical implications for each group of writers.

Finally, it is worth noting that for lexical diversity and cohesion, the comparison across all language groups of writers presented different results from the generic comparison between native and nonnative writers. Thus, it supports the diversity of NNSs' lexical performance in writing, which is in line with Altenberg and Granger (2001), Chrabaszcz and Jiang (2014), and Paquot (2010). The characteristics of different writers and the presented significant differences between writers from diverse mother tongue groups appeal for acknowledgement of learner linguistic diversity and the essentiality of targeted and tailored vocabulary instruction in academic writing.

Significance of the Findings

The primary contribution of this study is that it offers a systematic and thorough examination of the lexical features in native and nonnative English speakers' argumentative writing. Because this study used corpus data instead of surveys or questionnaires (Ostler, 1980), the descriptions are reliable, unbiased, and generalizable (McEnery, Xiao & Tono, 2006). Subsequently, through statistical analyses conducted on top of the descriptive features of the texts, the results identified statistically significant differences in lexical performance between native and nonnative writers as well as across all writers from different language backgrounds.

The differences shed light on the diverse needs of NNSs in academic English writing, particularly in the lexical aspect.

Although the results of the present study are different from some of the previous studies because of different population and approaches to measure the lexical features, the current study adds directly to renowned research in the field of using Corpus Linguistics approaches to study vocabulary and writing. The deviations between the findings confirm the diversity in TESOL and that one size really does not fit all.

The pedagogical significance of the present study can be shown in the following two scenarios. First, imagine in a college composition class in the U.S. where there might be 18 international students from six different countries, including China, Germany, Japan, Mexico, Russia, and Turkey. Oftentimes, intro-level composition courses are taught by graduate students or novice instructors. Thus, it is more than likely that these instructors may determine their international students' writing abilities and needs based on their general knowledge of English learners, which does not equip them to fully analyze the actual and diverse learner needs. As a result, appropriate and targeted lesson planning is highly unlikely to happen in this case. The findings of the current study can play an important role in assisting less experienced instructors to have a better idea of what to expect in their NNSs' writing. For instance, understanding that Turkish writers may lack the knowledge of employing academic and low-frequency vocabulary in their writing, instructors may realize that adding a list of commonly used academic words in the lesson plan can be beneficial for fulfilling Turkish students' needs in writing.

Second, a more likely classroom scenario is a group of international students with a couple of dominant first languages. In the state of Florida, having international students with Chinese and Spanish language backgrounds in a college classroom is very common. For

instructors in almost all discipline areas, spending time on learning those international students' mother tongues to better understand their needs seems infeasible and unpractical. However, with the findings from the current study, the instructors are better off analyzing their students' unique needs in writing, recognizing the challenges that the students face, fine tuning the lesson plans, and providing them with effective assistance to improve their academic writing performance.

Another significance of the findings lies in the aspect of material and curriculum designing. The results of the current study challenge some of the common assumptions of L1 influence based on historical linguistics. For instance, it is generally considered that German speakers might outperform other nonnative writers since both German and English belong to Germanic language. Similarly, Spanish writers might also benefit from the connection between Romance language and English with the common Latin origin. Yet, the results indicate that on average, German writers might actually face more problems with employing academic vocabulary than do other groups of NNSs. In addition, compared to other groups of nonnative writers, Spanish writers' performance did not display their L1 advantages in any of the three lexical features.

Hence, for material and curriculum designers, besides using common knowledge from historical linguistics, results of the present research demonstrate the importance of using empirical evidence to analyze learner characteristics and further conduct needs analysis. Understanding the differences between native and nonnative writers' lexical performance provides the foundation for compiling vocabulary teaching materials to teach academic writing. Meanwhile, acknowledging the diversity among NNSs is essential for developing complementary materials for specific groups of learners. Depending on the results of needs analysis in terms of learner characteristics, material writers can provide appendices with targeted

assignments or in-class activities to fulfill learners' special needs. For the six groups of NNSs in this study, their varied needs are analyzed in detail shortly.

In sum, the significance of the findings in the present study lies in providing empirical evidence for instructors and material designers in various academic settings to realize what exactly NNSs might not be capable of performing when compared to NSs. Furthermore, the findings presented characteristics of NNSs from various language backgrounds, which indicates different demands of the NNSs. Thus, the findings call for targeted vocabulary instruction in academic writing and are beneficial for modifying instructional strategies according to learner needs. Specific recommendations and implications on how to modify the instruction for various groups of NNSs are presented shortly in the following section.

Limitation of the Study

A few limitations may apply to the present study. First, even though the foundation corpora (i.e., the ICLE and the LOCNESS) have been employed in various empirical studies and have been validated as reliable sources. In order to control the scope of the present study, only argumentative essays were selected to represent academic writing. Admittedly, argumentative writing is one of the most popular genres for assessment in university courses. However, academic writing surely involves a broader range of genres besides argumentative essays, such as narratives, reports, reviews, and so forth. Hence, merely employing argumentative essays to generalize the field of academic writing is one of the limitations for the current study.

Second, the present study neglected the influence of spelling errors and certain formatting issues in the essays. Adjusting spelling mistakes and formatting the essays require time-consuming manual checking, which could not be afforded in this study. Thus, the spelling errors and certain formatting issues might skew the results to some extent.

Lastly, the current study focuses on presenting quantitative measures of lexical features in the essays, which leaves the qualitative aspect of writing evaluation unattended but open for future research. The qualitative aspect could be helpful in revealing whether the vocabulary is correctly used. The sentence “I think this phenomena is really scary” is from a Japanese writer’s text, which contains the academic word *phenomena*. However, the writer did not write the correct singular form of the word, namely *phenomenon*. Thus, without correct usage, the quality of the writing could be reduced even with diverse or sophisticated vocabulary.

Pedagogical Implications

In this section, implications for instructional practice are provided in detail based upon the findings of the current study. To begin with, the findings of the first research question showed that compared to NSs, NNSs 1) had lower levels of lexical diversity; 2) significantly overused high-frequency words and underused academic and low-frequency words; 3) had more uses of word repetition and connectives but lower levels of semantic cohesion. Hence, in general, instructors could be advised to provide lists of academic vocabulary, low-frequency vocabulary, or discipline-specific vocabulary to NNSs. Improving NNSs vocabulary size and diversity also offers more options for them to build deeper-level cohesion in writing.

The second research question tackled the characteristics of each group of writers from various language backgrounds and revealed significant differences between them in lexical performance. In the following paragraphs, some guidelines for working with students from these individual language backgrounds are offered. Because of the similar or common challenges that some language groups face, some of the recommendations are analogous.

Chinese Writers. Among all six groups of NNSs, Chinese speakers’ essays ($n = 100$) presented the most similarities to those of NSs. Except for the significant differences from NSs

in lexical diversity and the coverages of the second 1,000 words, Chinese writers' texts had low coverage of high-frequency words and high coverage of academic and off-list words. In terms of cohesion, Chinese writers demonstrated higher levels of referential cohesion and LSA by using limited connectives. In addition, it is notable that Chinese writers' LSA was significantly higher than that of NSs.

As a result, one pedagogical implication for teaching Chinese writers can be focusing on diversifying the words that the writers use in production. This strategy can also be helpful for controlling Chinese writers' word level repetition (i.e., referential cohesion). This strategy should go beyond merely pointing out whether a paper lacks lexical diversity or not; rather, the instructors should provide their Chinese English learners with explicit substitutable vocabulary or synonyms of the commonly used words, advanced vocabulary lists, detailed feedback, and textual examples in how to practically achieve lexical diversity. Laufer (1994) also suggested creating lexical syllabi and integrating vocabulary teaching and practicing to the existing curriculum.

German Writers. German writers' texts ($n = 97$) displayed the highest level of lexical diversity and the highest coverage of the second most frequent 1,000 words. However, German writers' coverages of the AWL and the AVL as well as the LSA level were the lowest among all NNSs. In terms of other lexical features, German writers' performance was in the medium range. This could be different from most instructors' expectation as both German and English belong to Germanic language. One might assume that compared to Chinese writers, whose mother tongue is much different from English, German writers might have less difficulties in academic English writing. Nevertheless, the empirical results of the current study overturned the assumption and

indicated that the focus for instructing German writers should be improving their knowledge of academic vocabulary and semantic cohesion.

To illustrate German writers' lexical performance more directly, the following paragraph is selected from a German speaker's writing:

It is also true that the children of today have a better education than in former times, when education was a privilege reserved to the rich. Our kids can all attend primary schools, secondary schools, grammar schools, comprehension schools and diving schools; they can go to university if they choose to become an academic, for education is free!

In this paragraph, there are 58 total running words. However, only two words are in the AWL band: *primary* and *academic*; only four words belong to the AVL: *primary*, *than*, *university*, and *academic*. There are some words in the text could be replaced by more advanced or academic vocabulary. For instance, the word *former* can be replaced by *previous*; *rich* can be substituted by *wealthy*, and so forth.

To implement the vocabulary instruction for German writers, various academic or discipline-specific vocabulary lists can be introduced to them. For instance, the AWL and the AVL that were employed in the current study are both freely available on the Internet. Both lists have been validated with high coverages of most academic texts (Coxhead, 2000; Gardner & Davies, 2013). Currently, there has been a wide array of vocabulary lists accessible online, while introducing these lists to the NNSs, instructors are advised to understand how the words are selected and if the list is representative.

Meanwhile, more instruction should be given to NNSs in building semantic coreferentiality in their writing. For instance, teaching groups of vocabulary based on their

shared thematic concepts appears to be beneficial for facilitating vocabulary learning (Folse, 2004). Having the receptive knowledge of vocabulary with connected thematic notions is one possible strategy for improving productive knowledge in writing.

Japanese Writers. Findings of Japanese writers' essays ($n = 97$) demonstrated that Japanese writers performed poorly in lexical diversity and lexical sophistication. It was reflected with the lowest level of lexical diversity, high coverage of the high-frequency words, the second lowest coverage of the academic words, and relatively low coverage of low-frequency words. In terms of cohesion, Japanese writers' texts revealed relatively high levels of referential cohesion, LSA, and frequent use of connectives. The following text is extracted from a Japanese writer's essay:

I think the greatest invention of the twentieth century is the Internet. It is used by all over the world now. I use it every day. I gather information about the Waseda University, professional baseball and so on. I check classes information and I am able to know whether today's classes are held or not.

In this fragment, the number of the total running words is 55, however, the word *I* was used for five times, *it* was used twice, *information* was used twice, *classes* was used twice. In addition, the coverage of the first 2,000 words is above 92% in this text. The only AWL word in this text is *professional*. To improve the lexical performance of this specific text, instructors can teach students some formulaic expressions that can help them diversify the vocabulary. For instance, *I think* can be substituted by the phrase *in my opinion*. Moreover, the short sentence *I use it every day* is not a good example of academic expression; instead, the student can write it as *The Internet is beneficial for my daily life*, in which *beneficial* is an AWL word.

Consequently, the pedagogical strategies for instructing Japanese English writers should focus on improving their vocabulary size and sophistication level as well as diversifying their use of vocabulary in writing. With a larger vocabulary size, it is more possible for the writers to vary their lexis in writing. Specific approaches can be found in the previous implications for Chinese and German writers, which may include providing vocabulary lists, substitutional words, synonyms, and exemplary writing.

Russian Writers. Russian writers' ($n = 96$) lexical diversity level was the second highest. However, most of the measures showed a medium level of lexical performance in lexical sophistication. Regarding Russian writers' performance in cohesion, the referential cohesion and the employment of connectives were the lowest among all seven groups of writers; in addition, the LSA level was relatively low as well. The follow paragraph is from a Russian writer's essay:

It is really true that the world of science and technology is rapidly developing. Every year life in the human society becomes more civilised. New technologies, new machines, new services are invented for the people, our modern civilisation spreads. New discoveries make the standards of living higher and higher. Our modern world is completely rational and science, computers and machines really play an extremely important part in our daily lives. We have such great services as electricity, television, radio and many others at our disposal. Sometimes it even seems that there is nothing more to discover.

Despite some grammatical errors, this paragraph lists several examples of the development of science and technology. Logically, the sentences in the paragraph are in a coordinative relationship. Nevertheless, no cohesive devices are used to connect the sentences. To improve, instructors can introduce some transition words or phrases that can be used to connect the coordinative relationship, such as *in addition*; *moreover*, *furthermore*, and so forth.

Therefore, the focus for instructing Russian writers should be on improving their lexical performance in building cohesion in writing. The initial step could include introducing transitional vocabulary for building different logical connections between the sentences. Moreover, templates for structuring formal academic sentences and texts as well as exemplary samples should be provided for the writers. Explicit instruction on how to connect sentences and build relationships between sentences should be offered as well.

Spanish Writers. Spanish writers' lexical features in their texts ($n = 99$) were not prominent in either category. Most of the results lay in the modest levels of performance. Therefore, besides regular instruction on vocabulary and academic writing, it is advised to evaluate Spanish writers' lexical quality individually to obtain more detailed needs of the learners based on the circumstances. Many of the forementioned approaches for other NNSs can be employed as well.

Turkish Writers. Not significantly different from Japanese writers, Turkish writers' performance regarding lexical diversity and lexical sophistication in their texts ($n = 95$) was among the lowest in NNSs. Their coverage of the high-frequency words, namely the first and second 1,000 words, was the highest. This statistic therefore leads to the consequence that the coverages of low-frequency words and academic words in Turkish writer's texts were exceedingly low. In terms of cohesion, even though Turkish writers' use of connectives was the highest, their levels of referential cohesion and the LSA were relatively low. For instance, the following paragraph is extracted from a Turkish writer's essay:

First of all I want to say that I don't believe sex equality. Of course, there should be equality between them but in real time there isn't because their duties are different. For example, lets talk about men they are fathers, brothers, husbands and their duties are

earning money protecting the home and providing the necessities. So they should be strong.

In this paragraph, there are 61 running words in total; however, only one word, *sex*, belongs to the AWL and only five words (i.e., *between*, *example*, *providing*, *equality*, and *necessities*) belong to the AVL. The coverage of the most frequent 2,000 words is more than 96%. The MTLD value that was used to measure lexical diversity is 54.88 in this text, which is below the average score 70 to 120 (McCarthy & Jarvis, 2010). In terms of cohesion, indeed several transition words and phrases were employed in this text, such as *first of all*, *for example*, and *so*. Nevertheless, there is also apparent disconnection between the sentences. For instance, in the second sentence, the pronoun *them* does not clearly refer to any of the words in the previous sentence. As a result, in addition to the approaches that are mentioned in prior groups of writers to improve learners' lexical diversity and lexical sophistication levels, instructors should also help Turkish writers focus on building deeper level cohesion in their writing.

In sum, pedagogical strategies for improving Turkish writers' lexical performance should include all three lexical features. For lexical diversity and lexical sophistication, as suggested for other groups of writers, vocabulary lists of academic, substitutional, and discipline-based words should be provided. For improving lexical performance in cohesion, since Turkish writers presented the existing knowledge of connectives, more instruction can be focused on how to appropriately use these connectives to build more logical connections between the sentences and the paragraphs in the texts.

To conclude, NNSs' lexical performance displayed varying needs for adjusting pedagogical strategies in teaching vocabulary and academic writing. Previous research has revealed the positive impacts of explicit and targeted vocabulary instruction on improving NNSs'

lexical performance in academic writing (Lee, 2003; Young-Davy, 2014). NNSs who are in need of supports regarding lexical diversity and lexical sophistication are German, Japanese, and Turkish writers. Providing synonyms and other substitutable vocabulary can be an effective way for diversifying their use of vocabulary in writing. In addition, lists of academic, technical, and discipline-specific vocabulary can be offered to enhance their lexical sophistication level. NNSs who demonstrated the demand of improving lexical performance in cohesion are Russian and Turkish writers. Explicit instruction and feedback on how to establish lexical and semantic coreferentiality would be beneficial (Harman, 2013). Instructors should give specific feedback on the lack of cohesion; in addition, they can perhaps provide NNSs with exemplary texts and underline how those texts use appropriate connectives and lexical networks to build semantic cohesion.

Recommendations for Further Research

The results of the current study can serve as the foundation for further exploring the characteristics of NNSs' academic writing and what they truly need to improve their writing performance. Recommendations for further research to deepen the understanding in lexis and L2 writing can include expanding both the width and depth of the present study.

To broaden the scope of the present study, suggestions for further study would be recruiting NNSs from more diverse language backgrounds, examining other genres of academic writing, and including other lexical features for analysis. The current study only extracted essays of NNSs from six mother tongue groups, which is far from sufficient to cover the diversity of NNSs. With the increase of international students from various countries, it is necessary to enlarge the range of NNSs' language backgrounds. Moreover, different language backgrounds could refer to various educational environments and teaching approaches. In terms of other

genres of academic writing, it has been mentioned previously that academic writing includes other commonly assessed genres as well, such as reports, narratives, and reviews. Therefore, argumentative writing, which was examined in the present study, merely represents academic writing segmentally. Finally, some lexical features that have been studied in prior literature were not included in the present study due to practical difficulties of time and space. For example, lexical bundles and lexical errors are considered as important aspects for evaluating the writing quality, however, they were not included in the present study. Hence, evaluating other essential lexical features of the texts can be beneficial for revealing more demands of the NNSs in academic writing.

With respect to further the depth of the present study, analyzing the texts from a qualitative perspective could be one promising aspect for the future research. Analyzing whether sophisticated vocabulary was correctly and appropriately used in the texts might reveal deeper quality of the writing. Some computational tools for annotating the texts could be helpful for detecting the usages of the vocabulary. Qualitative approaches can be also used to investigate the educational backgrounds of different groups of nonnative writers. Understanding their educational environment of English learning could be inspirational for further analyzing the reasons that caused the diversity in their lexical performance. Interviews, journals, observations, and surveys with open ended questions are some commonly used approaches for exploring the insights of this issue. Furthermore, it is critical to note that the current study focused on concluding the lexical performance of different groups of writers by analyzing their average performance, meaning the individual differences in each group were not emphasized. However, the individual differences within each group of writers do exist. Hence, conducting a more

detailed analysis in terms of individual differences could provide more valuable descriptions of the learner characteristics and more precise pedagogical advice.

Conclusion

The present study examined the characteristics of native and nonnative English writers' lexical performance in academic writing, focusing on three lexical features of lexical diversity, lexical sophistication, and cohesion. These comparisons in lexical features were conducted both between native and nonnative English writers as well as across all writers from the six language backgrounds selected for inclusion in this study.

Quantitative analyses revealed differences between native and nonnative English writers' performance in all three major lexical features. The results demonstrated that NNSs failed to reach the same levels as NSs in lexical diversity, lexical sophistication, and cohesion. In particular, the differences between native and nonnative writers in lexical sophistication were statistically significant.

The findings of the comparison across all writers from various language backgrounds revealed statistically significant differences between the NNSs in all three lexical features, which suggested the diversity of NNSs and the varied learner needs in improving lexical performance in L2 academic English writing.

APPENDIX A: IRB LETTER OF APPROVAL



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

NOT HUMAN RESEARCH DETERMINATION

From: UCF Institutional Review Board #1
FWA00000351, IRB00001138
To: Xiaoli Yu
Date: November 22, 2017

Dear Researcher:

On 11/22/2017, the IRB determined that the following proposed activity is not human research as defined by DHHS regulations at 45 CFR 46 or FDA regulations at 21 CFR 312.61:

Type of Review: Not Human Research Determination
Project Title: Two-level Comparisons of Lexical Features in Academic
Writing between Native and Nonnative Writers and across
Nonnative Writers
Investigator: Xiaoli Yu
IRB ID: SBE-17-13545
Funding Agency:
Grant Title:
Research ID: N/A

University of Central Florida IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are to be made and there are questions about whether these activities are research involving human subjects, please contact the IRB office to discuss the proposed changes.

This letter is signed by:

A handwritten signature in black ink, appearing to read "Gillian Morien".

Signature applied by Gillian Morien on 11/22/2017 11:37:54 AM EST

Designated Reviewer

APPENDIX B: COPYRIGHT PERMISSION LETTER

5/30/2018

Mail - Xiaoli Yu@ucf.edu

RE: Reprint Permission Request

Sylviane Granger <sylviane.granger@uclouvain.be>

Wed 5/30/2018 12:21 PM

To: Xiaoli Yu <Xiaoli.Yu@ucf.edu>;

1 attachments (308 KB)

Granger_CIA2_JLCR2015.pdf;

Dear Xiaoli,

I'm willing to grant you permission to reproduce the figure provided you include the exact references to the original article.

I suppose you are aware of the fact that I revisited the CIA framework in 2015. I attach a copy of the paper for your information.

Best wishes,

Sylviane Granger

Professor Sylviane Granger
Centre for English Corpus Linguistics
Université catholique de Louvain
Place Blaise Pascal 1
B-1348 Louvain-la-Neuve (Belgium)

<https://ucf.edu/offices-services/translation/>

10

5/30/2018

Mail - Xiaoli Yu@ucf.edu

<https://uclouvain.be/en/research-institutes/ilc/cec/>
<https://uclouvain.be/en/directories/sylviane.granger>
https://www.researchgate.net/profile/Sylviane_Granger
<http://www.lernercorpusassociation.org/>

De : Xiaoli Yu [mailto:Xiaoli.Yu@ucf.edu]

Envoyé : mercredi 30 mai 2018 17:14

À : Sylviane Granger

Objet : Reprint Permission Request

Dear Professor Granger,

I am a doctoral student in TESOL at the University of Central Florida. I am writing to kindly ask for your permission to use one figure from one of your book chapters in my doctoral dissertation.

Following is the information of the book chapter that I extracted the figure from:

Book chapter: From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora.

Page: 44 Figure 2 (CIA structure figure)

Book: Languages in contrast: Papers from a symposium on text-based cross-linguistic studies

My dissertation is titled as:

One size does not fit all: Lexical differences between native and nonnative academic English writing and variability among nonnative English writers

<https://ucf.edu/offices-services/translation/>

20

5/30/2018

Mail - Xiaoli Yu@ucf.edu

I am attaching the abstract of my dissertation here for your reference as well.

Thank you very much for your time and consideration.

Xiaoli Yu

Graduate Teaching Assistant & Doctoral Candidate in TESOL
College of Education and Human Performance
University of Central Florida

<https://ucf.edu/offices-services/translation/>

20

APPENDIX C: COPYRIGHT PERMISSION LETTER

07/2018

Mel - Xiaoli Yu@ucf.edu

RE: Reprint Permission Request

Sylviane Granger <sylviane.granger@uclouvain.be>

Fri 6/1/2018 3:30 AM

To: Xiaoli Yu <Xiaoli.Yu@ucf.edu>;

Dear Xiaoli,

You can reproduce the three figures provided that you include the full reference (authors, date, page number) in the caption of each figure. I would also appreciate if you could send me a copy of your thesis once you have completed it.

Best wishes,
Sylviane Granger

De : Xiaoli Yu [mailto:Xiaoli.Yu@ucf.edu]

Envoyé : jeudi 31 mai 2018 16:27

À : Sylviane Granger

Objet : Re: Reprint Permission Request

Thank you very much Dr. Granger.

Following are the tables from the ICLE V2 handbook that I would like to use in my dissertation:

Source: Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). International corpus of learner English V2. Presses Universitaires

<https://icbale.office.com/news/7readmucf.edu>

10

07/2018

Mel - Xiaoli Yu@ucf.edu

de Louvain.

Table 3: Top ten essay topics (pp. 6)

Table 6: CEF results-20 essays per subcorpus (pp. 12)

Table 7: Distribution of essays/words per subcorpus (pp. 25)

I appreciate your time and support.

Best,

Xiaoli Yu

From: Sylviane Granger <sylviane.granger@uclouvain.be>

Sent: Wednesday, May 30, 2018 3:04:44 PM

To: Xiaoli Yu

Subject: RE: Reprint Permission Request

Dear Xiaoli Yu,

I do have another question. Regarding the copyright of the handbook ICLE V2, do you know if I should contact Presses universitaires de Louvain for their permission to use a few tables in the book?

You can send the request directly to me.

Best wishes,

Sylviane Granger

<https://icbale.office.com/news/7readmucf.edu>

20

APPENDIX D: COPYRIGHT PERMISSION LETTER

6/11/2018

Mail - Xiaoli.Yu@ucf.edu

Fwd: Permission Request Form

Rights - Ineke Elskamp <rights@benjamins.nl>

Tue 6/5/2018 5:50 AM

To: Xiaoli Yu <Xiaoli.Yu@ucf.edu>;

 1 attachments (415 KB)

permission request.pdf;

Dear Xiaoli Yu,

Herewith we have the pleasure to grant you permission to use Figure 2 as published on page 17 of:

"Contrastive interlanguage analysis. A reappraisal", Granger, S, in our journal: "International Journal of Learner Corpus Research" - Volume 1:1 - 2015.

John Benjamins Publishing Company. Amsterdam/Philadelphia.

The Figure will be reprinted in your upcoming doctoral dissertation with the working title: "Lexical differences between native and nonnative academic English writing and variability among nonnative English writers"

To be published by the: University of Central Florida database.

Estimated publication date: August 2018

Please note that this permission is given on condition that full acknowledgement of the original source is given.

A link to the book on our website is fine as well: <https://benjamins.com/catalog/ijlcr>

Kind regards,
Ineke

--

[Subscribe to our monthly newsletter](#)

LIST OF REFERENCES

- Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81-92. doi: 10.1016/j.esp.2011.08.004
- Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-77). Amsterdam: John Benjamins Publishing Company.
- Allen, M. (2017). *The sage encyclopedia of communication research methods* (vols. 1-4). Thousand oaks, CA: SAGE publications.
- Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80-93). New York, NY: Addison Wesley Longman Limited.
- Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22(2), 173-195. doi: 10.1093/applin/22.2.173
- Anderwald, L., & Wagner, S. (2007). FRED—The Freiburg English dialect corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora* (pp. 35-53). London, UK: Palgrave Macmillan.
- Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic English: Implications for K-12 English language learners*. Retrieved from The George Washington University, Center for Equity and Excellence in

Education website:

<https://pdfs.semanticscholar.org/76b8/476fd601e434e53b6c6edd2855b5e2fe1b45.pdf>

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Bailey, A. L. (2007). *The language demands school: Putting academic English to the test*. New Haven, CT: Yale University Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Report 611). Retrieved from University of California, Los Angeles, Center for the Study of Evaluation website: <https://files.eric.ed.gov/fulltext/ED483026.pdf>

Bailey, A. L., Butler, F. A., & Sato, E. (2005). *Standards-to-standards linkage under title III: Exploring common language demands in ELD and science standards* (CSE Technical Report 667). Retrieved from University of California, Los Angeles, Center for the Study of Evaluation website: <https://files.eric.ed.gov/fulltext/ED492890.pdf>

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en François L2. *Journal of French Language Studies*, 14(3), 281-299. doi: 10.1017/S0959269504001802

Belz, J., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: The case of German modal particles. *Canadian Modern Language Review*, 62(1), 17-48. doi: 10.3138/cmlr.62.1.17

Berman, R. (1994). Learners' transfer of writing skills between languages. *TESL Canada Journal*, 12(1), 29-46. doi: 10.18806/tesl.v12i1.642

- Biber, D., & Reppen, R. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge, UK: Cambridge University Press.
- Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis* (Research Report No. TOEFL iBT-19). Retrieved from Educational Testing Service (ETS) website:
<https://www.ets.org/Media/Research/pdf/RR-13-04.pdf>
- 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*. Essex, UK: Pearson Education Ltd.
- Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 548-570). New York: Oxford University Press, Inc.
- Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, 33(1), 1-17. doi:10.1111/j.1467-1770.1983.tb00983.x
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, 14(1), 14-31. doi: 10.1080/10904018.2000.10499033
- Brewer, E. W., & Kubn, J. (2012). Causal-comparative design. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 125-131). Thousand Oaks, CA: SAGE Publications, Inc.
- British Council. (2013). *The English effect*. Retrieved from
<https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf>

- Britton, J. (1975). *The development of writing abilities (11-18)*. Urbana, IL: National Council of Teachers of English.
- Browne, C., Culligan, B., & Phillips, J. (2013). *The new general service list*. Retrieved from <http://www.newgeneralservicelist.org/>
- Brutt-Griffler, J., & Samimy, K. K. (2001). Transcending the nativeness paradigm. *World Englishes*, 20(1), 99-106. doi:10.1111/1467-971X.00199
- Burke, E., & Wyatt-Smith, C. (1996). Academic and non-academic difficulties: Perceptions of graduate non-English speaking background students. *TESL-EJ*, 2(1). Retrieved from <http://www.teslj.org/wordpress/issues/volume2/ej05/ej05a1/>
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide, spoken and written English, grammar and usage*. Cambridge, UK: Cambridge University Press.
- Castro, C. D. (2004). Cohesion and the social construction of meaning in the essays of Filipino college students writing in L2 English. *Asia Pacific Education Review*, 5(2), 215-225. doi: 10.1007/BF03024959
- Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3), 267-296. doi: 10.1016/S1060-3743(03)00038-9
- Chen, J. (2008). An investigation of EFL students' use of cohesive devices. *Asia Pacific Education Review*, 5(2), 215-225.
- Chen, Q., & Ge, C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, 26(4), 502-514. doi:10.1016/j.esp.2007.04.003

- Chrabaszcz, A., & Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: A cross-linguistic comparison. *Second Language Research*, 30(3), 351-379. doi:10.1177/0267658313493432
- Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185-206). Cambridge, UK: Cambridge University Press.
- Cobb, T. (n.d.). Compleat lexical tutor. Retrieved from <http://www.lex tutor.ca/>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Prentice-Hall.
- Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33(4), 497-505. doi: 10.1080/14640748108400805
- Corrigan, K., & Buchstaller, I. (2007). *Handbook to the Newcastle electronic corpus of Tyneside English 2 (NECTE2)*. Retrieved from <https://research.ncl.ac.uk/necte2/documents/handbook.pdf>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238. doi: 10.2307/3587951
- 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., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16(3), 129-147. doi: 10.1016/j.jslw.2007.07.002

- Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2), 119-135. doi: 10.1016/j.jslw.2009.02.002
- Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3), 170-191. doi: 10.1504/IJCEELL.2011.040197
- Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3), 475-493. doi: 10.1002/j.1545-7249.2008.tb00142.x
- Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15-30. doi: 10.1111/j.1540-4781.2007.00507.x
- Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, 45(1), 182-193. doi: 10.5054/tq.2010.244019
- Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the Cognitive Science Society, USA*, 32(32), 984-989.
- Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question, and some other matters. *Working Papers on Bilingualism*, 19, 121-129.
- Davies, M. (2008-). The corpus of contemporary American English (COCA): 560 million words, 1990-present. Retrieved from <http://corpus.byu.edu/coca/>

- De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 51-68). Amsterdam: Rodopi B.V.
- De Haan, P., & Van Esch, K. (2005). The development of writing in English and Spanish as foreign languages. *Assessing Writing*, 10(2), 100-116. doi: 10.1016/j.asw.2005.05.003
- Díez-Bedmar, M. B., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, P. Papp & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 147-175). Amsterdam: Rodopi.
- Douglas, S. R. (2010). *Non-native English speaking students at university: Lexical richness and academic success* (Doctoral dissertation). Retrieved from University of Calgary Archives.
- Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, 25(2), 220-242. doi: 10.1093/applin/25.2.220
- Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47(2), 157-177. doi: 10.1515/iral.2009.007
- Eckstein, G., & Ferris, D. (2018). Comparing L1 and L2 texts and writers in first-year composition. *TESOL Quarterly*, 52(1), 137-162. doi: 10.1002/tesq.376
- Eia, A. B. (2006). *The use of linking adverbials in Norwegian advanced learners' written English* (Master's thesis). Retrieved from <http://urn.nb.no/URN:NBN:no-16093>
- Ellis, R. (2008). *The study of second language acquisition*. New York: Oxford University Press.
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139-155. doi: 10.1016/1060-3743(95)90004-7

- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2014). G*Power Version 3.1.2 [Computer software]. Universität Kiel, Kiel, Germany. Retrieved from <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register>
- Ferguson, G. (2007). The global spread of English, scientific communication and ESP: Questions of equity, access and domain loss. *Ibérica, Revista De La Asociación Europea De Lenguas Para Fines Específicos (AELFE)*, 13, 7-38.
- Fernsten, L. A., & Reda, M. (2011). Helping students meet the challenges of academic writing. *Teaching in Higher Education*, 16(2), 171-182. doi: 10.1080/13562517.2010.507306
- Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, 28(2), 414-420. doi: 10.2307/3587446
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.
- Field, Y., & Oi, Y. L. M. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal*, 23(1), 15-28. doi: 10.1177/003368829202300102
- Flowerdew, J. (2006). Use of signalling nouns in a learner corpus. *International Journal of Corpus Linguistics*, 11(3), 345-362. doi: 10.1075/ijcl.11.3.07flo
- Flowerdew, J. (2010). Use of signalling nouns across L1 and L2 writer corpora. *International Journal of Corpus Linguistics*, 15(1), 36-55. doi: 10.1075/ijcl.15.1.02flo
- Flowerdew, L. (1998). Integrating 'expert' and 'interlanguage' computer corpora findings on causality: Discoveries for teachers and students. *English for Specific Purposes*, 17(4), 329-345. doi: 10.1016/S0889-4906(97)00014-8
- Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press.

- Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin Company.
- Francis, W. N., & Kučera, H. (1979). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University, Department of Linguistics.
- French, A., Macedo, M., Poulsen, J., Waterson, T., & Yu, A. (2008). Multivariate analysis of variance (MANOVA) [PDF file]. Retrieved from <http://userwww.sfsu.edu/efc/classes/biol710/manova/MANOVAnewest.pdf>
- Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, 23, 1-16. doi: 10.1016/j.jslw.2013.10.001
- Gardner, D. (2013). *Exploring vocabulary: Language in action*. New York, NY: Routledge.
- Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327. doi: 10.1093/applin/amt015
- Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 418-436). Cambridge, UK: Cambridge University Press.
- Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, 1(1), 41-61. doi:10.1075/etc.1.1.05gil
- Gilquin, G., de Cock, S., & Granger, S. (2010). *Louvain international database of spoken English interlanguage*. Louvain-la-Neuve: Presses Universitaires de Louvain.

- Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-335. doi: 10.1016/j.jeap.2007.09.007
- Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19(2), 115-135. doi: 10.1016/S0889-4906(98)00015-5
- Gonzalez, M. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker academic compositions* (Doctoral dissertation). Retrieved from Electronic theses and dissertations.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2), 193-202. doi: 10.3758/BF03195564
- Granger, S. (1993). The international corpus of learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 57-69). Amsterdam: Rodopi B. V.
- Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today*, 10(3), 25-33. doi: 10.1017/S0266078400007665
- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145-160). Oxford, UK: Clarendon Press.

- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Philadelphia, PA.: John Benjamins Publishing Company.
- Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546. doi: 10.2307/3588404
- Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam and Philadelphia: John Benjamins.
doi:10.1075/scl.33.04gra
- Granger, S. (2012). Learner corpora. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 3235–3242). Oxford, UK: Wiley-Blackwell.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24. doi: 10.1075/ijlcr.1.1.01gra
- Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1), 17-27. doi: 10.1111/j.1467-971X.1996.tb00089.x
- Granger, S., & Wynne, M. (2000). Optimising Measures of Lexical Variation in EFL Learner Corpora. In J. M. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English: Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora* (ICAME 1998; pp. 249-259). Amsterdam: Rodopi B. V.

- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2002). *The international corpus of learner English handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge, UK: Cambridge University Press.
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International corpus of learner English*. Belgium: Presses universitaires de Louvain.
- Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123-145. doi: 10.1016/S1060-3743(00)00019-9
- Grant, N., & Fabrigar, L. (2007). Exploratory factor analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design*. Thousand Oaks, CA: SAGE Publications, Inc.
- Gravetter, F. J., & Wallnau, L. B. (2014). Introduction to the t statistic. In F. J. Gravetter & L. B. Wallnau (Eds.), *Essentials of Statistics for the Behavioral Sciences* (pp. 267-299). Boston, MA: Cengage Learning.
- Green, C. (2012). A computational investigation of cohesion and lexical network density in L2 writing. *English Language Teaching*, 5(8), 57-69.
- Greenbaum, S. (1988-). The international corpus of English. Retrieved from <http://www.ucl.ac.uk/english-usage/projects/ice.htm>
- Hall, C., Lewis, G. A., McCarthy, P. M., Lee, D. S., & McNamara, D. S. (2007). A Coh-Metrix assessment of American and English/Welsh legal English. *Coyote papers: Psycholinguistic and computational perspectives. University of Arizona Working Papers in Linguistics*, 15, 40-54.

- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Harman, R. (2013). Literary intertextuality in genre-based pedagogies: Building lexical cohesion in fifth-grade L2 writing. *Journal of Second Language Writing*, 22(2), 125-140. doi: 10.1016/j.jslw.2013.03.006
- Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2), 153-166. doi: 10.1080/09571730802389983
- Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning*, 12(2), 111-132.
- Hong, H., & Cao, F. (2014). Interactional metadiscourse in young EFL learner writing: A corpus-based study. *International Journal of Corpus Linguistics*, 19(2), 201-224. doi: 10.1075/ijcl.19.2.03hon
- Human subject regulation decision charts. (2016, February 16). Retrieved from <https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/index.html>
- Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publication.
- Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, 16(3), 148-164. doi: 10.1016/j.jslw.2007.07.005
- Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2), 235-253. doi: 10.1002/j.1545-7249.2007.tb00058.x
- Ishikawa, S. (2011). A new horizon in learner corpora studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpol (Eds.), *Corpora and language technologies in teaching, learning, and researching* (pp. 3-11). Glasgow, UK: University of Strathclyde Press.

- Ishikawa, S. I. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1, 91-118.
- Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE corpus: Exploiting the language learner's speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2), 119-125.
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), 57-84. doi: 10.1191/0265532202lt220oa
- Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13-43). Philadelphia, PA: John Benjamins North America.
- Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar: Based on the LOB corpus*. Oxford, UK: Clarendon Press.
- Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293-313). Cambridge, UK: Cambridge University Press.
- Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, 7(3), 700-715. doi: 10.1002/tesj.238
- Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning and Technology*, 5(3), 77-90.
- Kice, A. R. (2014, September 11). How U.S. colleges gauge international students' English skills. *U.S. News*. Retrieved from <https://www.usnews.com/education/blogs/international-student-counsel/2014/09/11/how-us-colleges-gauge-international-students-english-skills>

- Kintsch, W., & van Dijk, T. A. (1978). Cognitive psychology and discourse: Recalling and summarizing stories. In W. U. Dressler (Ed.), *Current Trends in Text Linguistics* (pp. 61-80). Berlin, Germany: Walter de Gruyter.
- Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, 4(5), 900-913. doi:10.4304/jltr.4.5.900-913
- Konstantakis, N. (2007). Creating a business word for teaching business English. *Elia: Estudios De Lingüística Inglesa Aplicada*, 7, 79-102.
- Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20(2), 148-161. doi: 10.1016/j.jslw.2011.02.001
- Kortmann, B., & Wagner, S. (2005). The Freiburg English dialect project and corpus. In B. Kortmann, T. Herrmann, L. Pietsch, & S. Wagner (Eds.), *A comparative grammar of British English dialects: Agreement, gender, relative clauses* (pp. 1-20). Berlin, Germany: Mouton de Gruyter.
- Kwon, S. (2009). Lexical richness in L2 writing: How much vocabulary do L2 learners need to use. *English Teaching*, 64(3), 155-174.
- Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34, 12-24. doi: 10.1016/j.jslw.2016.10.003
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757-786. doi: 10.1002/tesq.194

- Lardiere, D. (2003). Revisiting the comparative fallacy: A reply to Lakshmanan and Selinker, 2001. *Second Language Research*, 19(2), 129-143. doi:10.1191/0267658303sr216oa
- Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal*, 75(4), 440-448. doi: 10.1111/j.1540-4781.1991.tb05380.x
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London, UK: Palgrave Macmillan,
- Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, 25(2), 21-33. doi: 10.1177/003368829402500202
- Laufer, B. (2003). The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression. In V. Cook (Ed.), *Effects of the second language on the first* (pp. 19-31). Clevedon, UK: Multilingual Matters LTD.
- Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied Linguistics*, 26(4), 582-588. doi: 10.1093/applin/ami029
- Laufer, B. (2013). Vocabulary and writing. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-5). John Wiley and Sons, Inc. doi: 10.1002/9781405198431.wbeal1432
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-322. doi: 10.1093/applin/16.3.307
- Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647-672. doi: 10.1111/j.1467-9922.2010.00621.x
- Lee, S. H. (2003). ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, 31(4), 537-561. doi: 10.1016/j.system.2003.02.004

- Leech, G. (1997). *Teaching and language corpora: A convergence*. London: Longman.
- Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge, UK: Cambridge University Press.
- Leider, C. M., Proctor, C. P., Silverman, R. D., & Harring, J. R. (2013). Examining the role of vocabulary depth, cross-linguistic transfer, and types of reading measures on the reading comprehension of Latino bilinguals in elementary school. *Reading and Writing*, 26(9), 1459-1485. doi: 10.1007/s11145-013-9427-6
- Leki, I. (2017). *Undergraduates in a second language: Challenges and complexities of academic literacy development*. New York, NY: Routledge.
- Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. New York, NY: Routledge.
- Li, Y., & Qian, D. D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System*, 38(3), 402-411. doi:10.1016/j.system.2010.06.015.
- Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, 33(4), 623-636. doi: 10.1016/j.system.2005.02.002
- Lourdunathan, J., & Menon, S. (2017). Developing speaking skills through interaction strategy training. *The English Teacher*, 34, 1-18.
- Louwerse, M. M. (2004). A concise model of cohesion in text and coherence in comprehension. *Revista Signos*, 37(56), 41-58. doi: 10.4067/S0718-09342004005600004
- Ma, Y., & Wang, B. (2016). A Corpus-based Study of Connectors in Student Writing: A Comparison between a Native Speaker (NS) Corpus and a Non-native Speaker (NNS) Learner Corpus. *International Journal of Applied Linguistics and English Literature*, 5(1), 113-118. doi: 10.7575/aiac.ijalel.v.5n.1p.113

- Marco, M. J. L. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes*, 19(1), 63-86. doi: 10.1016/S0889-4906(98)00013-1
- Martinez, R., & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics*, 33(3), 299-320. doi: 10.1093/applin/ams010
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, 1-13. doi: 10.1016/j.system.2015.04.015
- McCarthy, M., & O’Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3-13). New York, NY: Routledge.
- McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone Student’s book 2a with audio CD/CD-rom*. Cambridge: Cambridge University Press.
- McCarthy, P. M., & Jarvis, S. (2007). Vocab: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459-488. doi: 10.1177/0265532207080767
- McCarthy, P. M., & Jarvis, S. (2010). MTLT, vocab-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392. doi: 10.3758/BRM.42.2.381
- McCarthy, P. M., Briner, S. W., Rus, V., & McNamara, D. S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao & S. R. Poteet (Eds.), *Natural language processing and text mining*. London: Springer-Verlag.
- McEnery, A. M., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. New York: Cambridge University Press.
- McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 364-380). New York, NY: Routledge.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.
- Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, 18(4), 393-407. doi: 10.1191/0267658302sr211xx
- Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, 26(1), 32-47. doi: 10.1093/applin/amh037
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235-244. doi: 10.1093/ijl/3.4.235
- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83-98). Bristol: Multilingual Matters.
- Muncie, J. (2002). Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System*, 30(2), 225-235. doi: 10.1016/S0346-251X(02)00006-4

- Muncie, J. (2002). Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System*, 30(2), 225-235. doi: 10.1016/S0346-251X(02)00006-4
- Nassaji, H. (2004). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Canadian Modern Language Review*, 61(1), 107-134. doi: 10.1111/j.1540-4781.2006.00431.x
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, 44(4), 529-539. doi: 10.1017/S0261444811000267
- Nation, I. S. P., & Webb, S. (2010). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge, UK: Cambridge University Press.
- National Center for Education Statistics. (2018). *English language learners in public schools*. Retrieved from https://nces.ed.gov/programs/coe/indicator_cgf.asp
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242. doi: 10.1093/applin/24.2.223
- Norment, N., Jr. (2002). Quantitative and qualitative analyses of textual cohesion in African American students' writing in narrative, argumentative, and expository modes. *CLA Journal*, 46(1), 98-132.

- O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, 12(2), 217-237. doi: 10.1177/026553229501200205
- Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, 26(1), 45-65. doi: 10.1007/s11145-012-9392-5
- Olsson, E. (2015). Progress in English academic vocabulary use in writing among CLIL and non-CLIL students in Sweden. *Moderna Språk*, 109(2), 51-74.
- Omidian, T., Beliaeva, N., Todd, L., & Siyanova-Chanturia, A. (2017). The use of academic words and formulae in L1 and L2 secondary school writing. *New Zealand Studies in Applied Linguistics*, 23(2), 39-59.
- Osborne, J. (2008). Adverb placement in post-intermediate learner English: A contrastive study of learner corpora. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 127-146). Amsterdam: Rodopi B. V.
- Ostler, S. E. (1980). A survey of academic needs for advanced ESL. *TESOL Quarterly*, 14(4), 489-502. doi: 10.2307/3586237
- Paquot, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 101-119). Amsterdam: John Benjamin Publishing Company.
- Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London, UK: Continuum.
- Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam & Philadelphia: John Benjamins Publishing.

- Pirttimaa, R., Takala, M., & Ladonlahti, T. (2015). Students in higher education with reading and writing difficulties. *Education Inquiry*, 6(1), 24277. doi: 10.3402/edui.v6.24277
- Polio, C. (2001). Research methodology in second language writing research: The case text-based studies. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 91-115). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Pop, A. (2016). Universities in English-speaking countries offering the most English-taught degrees. *Study Portals*. Retrieved from <http://www.mastersportal.eu/articles/1719/universities-in-english-speaking-countries-offering-the-most-english-taught-degrees.html>
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513-536. doi: 10.1111/1467-9922.00193
- Quirk, R. (1960). Towards a description of English usage. *Transactions of the Philological Society*, 59(1), 40-61. doi: 10.1111/j.1467-968X.1960.tb00308.x
- Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.
- Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing*, 1(2), 79-107. doi: 10.1016/1060-3743(92)90010-M
- Reppen, R., & Gordon, D. (2011). *Grammar and beyond level 2 student's book*. Cambridge: Cambridge University Press.
- Rissanen, M. (1993). The Helsinki corpus of English texts. In M. Kytö, M. Rissanen, & S. Wright (Eds.), *Corpora across the centuries: Proceedings of the first international colloquium on English diachronic corpora* (pp. 73-80). Amsterdam: Rodopi, B. V.

- Salkind, N. J. (2010). *Encyclopedia of research design*. Thousand Oaks, CA: SAGE Publications.
- Scarcella, R. (2002). Some key factors affecting English learners' development of advanced literacy. In M. J. Schleppegrell & M. C. Colombi (Eds.), *Developing advanced literacy in first and second languages: Meaning with power* (pp. 209-226). Mahwah, NJ: Lawrence Erlbaum.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Hampshire, UK: Palgrave Macmillan.
- Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145-171. doi: 10.2307/3588328
- Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27(4), 657-677. doi: 10.2307/3587400
- Sinclair, J. (1996). *EAGLES: Preliminary recommendations on corpus typology*. Retrieved from <http://www.ilc.cnr.it/EAGLES/corpus/corpus.html>
- Singhal, M. (2004). Academic writing and generation 1.5: Pedagogical goals and instructional issues in the college composition classroom. *The Reading Matrix*, 4(3), 1-13.
- Snellings, P., van Gelderen, A., & de Groot, K. (2004). The effects of enhanced lexical retrieval on second language writing: A classroom experiment. *Applied Psycholinguistics*, 25(2), 175-200. doi: 10.1017/S0142716404001092
- St John, E. (2001). A case for using parallel corpus and concordance for beginners of a foreign language. *Language Learning and Technology*, 5(3), 185-203.
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, 36(2), 139-152. doi: 10.1080/09571730802389975

- Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(04), 577-607. doi: 10.1017/S0272263109990039
- Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences* (5th ed.). New York, NY: Routledge.
- Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.
- Tan, M. (2005). Authentic language or language errors? Lessons from a learner corpus. *ELT Journal*, 59(2), 126-134. doi:10.1093/eltj/cci026
- Testa, J. (2016). *Journal selection process*. Retrieved from Clarivate Analytics website: <http://wokinfo.com/essays/journal-selection-process/>
- The CEFR levels. (2018). Retrieved from Council of Europe website: <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1), 77-101. doi: 10.1111/j.1540-4781.2012.01422.x
- UNESCO Institute for Statistics. (2016). *Global flow of tertiary-level students*. Retrieved from <http://uis.unesco.org/en/uis-student-flow>
- Upton, T., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20(4), 313-329. doi: 10.1016/S0889-4906(00)00022-3

- Valcourt, G., & Wells, L. (1999). *Mastery: A university word list reader*. Ann Arbor, Michigan: The University of Michigan Press.
- Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, 65, 139-150. doi: 10.1016/j.system.2016.12.013
- Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28(3), 170-182.
doi:10.1016/j.esp.2009.04.001.
- Webb, S. A. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, 65(3), 441-470. doi:
10.3138/cmlr.65.3.441
- Wimmer, G., & Altmann, G. (1999). Review article: On vocabulary richness. *Journal of Quantitative Linguistics*, 6(1), 1-9. doi: 10.1076/jqul.6.1.1.4148
- Witherell, S., & Department of State. (2016). *Open doors 2016 executive summary*. Retrieved from <https://www.iie.org/en/Why-IIE/Announcements/2016-11-14-Open-Doors-Executive-Summary>
- Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication*, 32(2), 189-204. doi: 10.2307/356693
- Woodford, K. (2001). *Cambridge learner's dictionary with CD-ROM*. Cambridge: Cambridge University Press.
- Writing across the curriculum. (2018). Retrieved from University of Central Florida website: <http://wac.cah.ucf.edu/>

- Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, 3(2), 215-229.
- Xue, X. (2015). *Ten-thousand English compositions of Chinese learners (the TECCL corpus), version 1.1*. The National Research Centre for Foreign Language Education, Beijing Foreign Studies University.
- Yang, H., & Wei, N. (2005). *College English learners' spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.
- Yonek, L. M. (2008). *The effects of rich vocabulary instruction on students' expository writing* (Doctoral dissertation). Retrieved from ProQuest.
- Young-Davy, B. (2014). Explicit vocabulary instruction. *ORTESOL Journal*, 31, 26-32.
- Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2), 236-259. doi: 10.1093/applin/amp024
- Zanettin, F. (1994). Parallel words: Designing a bilingual database for translation activities. In A. Wilson & T. McEnery (Eds.), *Corpora in language education and research: a selection of papers from Talc94* (Technical Papers, Volume 4; pp. 99-111). Lancaster, UK: UCREL.
- Zhang, L. J., & Anual, S. B. (2008). The role of vocabulary in reading comprehension: The case of secondary school students learning English in Singapore. *RELC Journal*, 39(1), 51-76. doi: 10.1177/0033688208091140
- Zhou, A. A. (2009). What adult ESL learners say about improving grammar and vocabulary in their writing for academic purposes. *Language Awareness*, 18(1), 31-46. doi: 10.1080/09658410802307923