



Corpus-based Analysis of Grammatical Complexity and Lexical Richness in Second Language Writing Performance

A thesis submitted in partial fulfillment of the requirements for master degree in
Teaching English as a Foreign Language

Supervisor:

M. Raouf Moini (Ph.D.)

Advisor:

Abbas Zare-ee (Ph.D.)

By:

Ali Mozaffari

University of Kashan

English Language Department

January, 2014

STATEMENT OF AUTHENTICATION

I, hereby, declare that the present work is the result of my own sincere and genuine effort and therefore is by no means plagiarized in any way.

Ali Mozaffari

ACKNOWLEDGMENTS

It falls to me to thank certain people for their help throughout my writing of this thesis. I cannot express enough gratitude to my encouraging supervisor, Dr. Moini for his continued support and encouragement. He offered his time to me so generously and willingly any time I needed his help and suggestion. I also would like to extend my sincere thanks to Dr. Zare-ee who primarily encouraged me to get myself involved in research. Without his comments, my thesis would not be completed. It is also incumbent upon me to thank my other good professors at the University of Kashan including Dr. Shams, Dr. Sami and Dr. Rahimi for providing me the learning opportunities. And the last but not the least, I would like to offer my heartfelt thanks to my parents who, like always, helped and supported me during the development of my thesis. To them, I dedicate this thesis.

ABSTRACT

There are many textual features in second language (L2) writing which contribute to the quality of writing. Grammatical complexity and lexical richness are two important text-related constructs which seem to be undeservedly neglected in second language writing. The existing literature suggests incongruous and even sometimes contradictory results about the role of these constructs in second language writing. The main objective of this study is to explore the relationship between grammatical complexity and the overall quality of second language writing as well as the relationship between lexical richness and the overall quality of second language writing. The current study employed a corpus of L2 texts written by Iranian EFL university students at the intermediate level. The corpus, named the Written English Corpus of Iranian Learners (WECIL), contained 350 L2 texts surpassing 180,000 running words. Unlike many previous studies, this study tried to incorporate several measures of grammatical complexity and lexical richness for its textual analyses. For grammatical complexity, fourteen measures were used to analyze the five types of grammatical complexity which were mean length of production unit, sentence complexity, subordination, coordination and particular structures. For the analysis of different aspects of lexical richness, lexical density, lexical diversity and lexical sophistication were used. As to the quality of L2 writing, two raters independently evaluated the texts of WECIL based on the holistic scoring guidelines of Test of Written English with the inter-rater reliability of .89. Afterwards, the texts of WECIL were divided into three levels of writing proficiency according to their assigned scores. As determined by one-way ANOVA tests, there were significant differences among three levels of writing proficiency for all measures of grammatical complexity. However, the Scheffe post-hoc analyses indicated that some measures of grammatical complexity were not found to show statistically significant differences across all three levels of writing proficiency. Also, no significant difference was found for lexical density, though lexical diversity and lexical sophistication showed statistically significant differences. Pearson's correlations suggested that subordination and sentence

complexity had the highest correlations with the quality of L2 writing ($r=.47$, $p<.000$; $r=.39$, $p<.000$) while lexical diversity and lexical sophistication were moderately correlated with the quality of L2 writing ($r=.50$, $p<.000$; $r=.47$, $p<.000$). In the light of these findings, students and teachers are encouraged to pay more attention to grammatical complexity and lexical richness in second language writing.

Key Words: Second Language Writing, Lexical Richness, Grammatical Complexity, Quality of Writing, Corpus

TABLE OF CONTENTS

STATEMENT OF AUTHENTICATION.....	II
ACKNOWLEDGMENTS.....	III
ABSTRACT.....	IV
LIST OF TABLES.....	VIII
LIST OF ABBREVIATIONS.....	X
CHAPTER ONE: INTRODUCTION.....	1
1.1. Background.....	1
1.1.1. Textual Features.....	3
1.1.2. Composing Processes.....	4
1.1.3. Writing Contexts.....	4
1.2. Complexity: a Complex Component.....	5
1.3. Grammatical Complexity.....	6
1.4. Lexical Richness.....	9
1.5. Statement of the Problem.....	10
1.6. Objectives of the Study.....	11
1.7. Research Questions and Hypotheses.....	12
1.8. Significance of the Study.....	13
1.9. Definition of Key terms.....	14
CHAPTER TWO: REVIEW OF THE RELATED LITERATURE.....	18
2.1. Overview.....	18
2.2. Modes of Writing.....	19
2.3. Methods of Assessing Writing.....	21
2.4. The Development of Language Development Measures.....	23
2.5. Fluency, Accuracy and Complexity.....	25
2.6. Methods of Calculation.....	26
2.7. Grammatical Complexity.....	27
2.7.1. Measures of Grammatical Complexity.....	29
2.7.2. Main Empirical Studies on Grammatical Complexity.....	32
2.8. Lexical Richness.....	41
2.8.1. Measures of Lexical Richness.....	43
2.8.2. Empirical Studies on Lexical Richness.....	45
CHAPTER THREE: METHODOLOGY.....	51
3.1. Overview.....	51
3.2. The Corpus of the Study.....	51

3.3. Rating Procedures.....	54
3.4. Measures of Grammatical Complexity.....	55
3.5. Measures of Lexical Richness.....	57
3.6. Data analysis.....	60
CHAPTER FOUR: RESULTS.....	61
4.1. Overview.....	61
4.3. Grammatical complexity and Quality of L2 Writing.....	62
4.3.1. Length of Production Unit.....	64
4.3.2. Sentence Complexity.....	69
4.3.5. Subordination.....	72
4.3.7. Coordination.....	77
4.3.9. Particular Structures.....	82
4.3.10. Summary of Results on Grammatical Complexity.....	87
4.4. Lexical richness and the Quality of L2 Writing.....	88
4.4.4. Summary of Results on Lexical Richness.....	92
CHAPTER FIVE: SUMMARY, DISCUSSION AND CONCLUSION.....	94
5.0. Overview.....	94
5.1. Summary of the Study.....	94
5.2. Discussion.....	95
5.2.1. Grammatical Complexity and the Quality of L2 Writing.....	96
5.2.2 Lexical richness and the quality of L2 writing.....	101
5.3. Conclusion.....	104
5.4. Implications of the Findings.....	105
5.5. Limitations of the Study.....	107
5.6. Suggestions for Future Research.....	107
References.....	109
Appendices.....	117

LIST OF TABLES

1.1. Second Language Writing Dimensions and Perspectives.....	2
1.2. Types and Measures of Grammatical Complexity.....	8
1.3. Measures of Lexical Richness.....	10
2.1. Main Text Types of Writing.....	20
2.2. Methods of Writing Assessment.....	22
2.3. Frequency-based Measures of Grammatical Complexity.....	30
2.4. Ratio-based Measures of Grammatical Complexity.....	31
2.5. Index-based Measures of Grammatical Complexity.....	32
2.6. Significant Measures of Grammatical Complexity.....	39
2.7. Measures of Lexical Richness.....	44
2.8. Significant Aspects of Lexical Richness.....	49
3.1. General Characteristics of the Participants.....	52
3.2. Written English Corpus of Iranian Learners.....	53
3.3. Holistic Scoring Rubric of TWE.....	54
3.4. Measures of Grammatical Complexity.....	56
3.5. Measures of Lexical Richness.....	59
4.1. Descriptive Statistics of Writing Assessment.....	62
4.2. Levels of Writing Proficiency.....	63
4.3. Descriptive Statistics of Length of Production Unit.....	64
4.4. ANOVA for Length of Production Unit.....	65
4.5. Scheffe Post-hoc Tests for Length of Production Unit.....	66
4.6. Correlations between Length of Production Unit and the Quality of L2 Writing.....	68

4.7. Descriptive Statistics of Sentence Complexity.....	70
4.8. ANOVA for Sentence Complexity.....	70
4.9. Scheffe Post-hoc Tests for Sentence Complexity.....	71
4.10. Correlation between Sentence Complexity and the Quality of L2 Writing.....	71
4.11. Descriptive Statistics of Subordination.....	72
4.12. ANOVA for Subordination.....	74
4.13. Scheffe Post-hoc Tests for Subordination.....	74
4.14. Correlations between Subordination and the Quality of L2 Writing.....	76
4.15. Descriptive Statistics for Coordination.....	78
4.16. ANOVA for Coordination.....	78
4.17. Scheffe Post-hoc Tests for Coordination.....	80
4.18. Correlations between Coordination and the Quality of L2 Writing.....	81
4.19. Descriptive Statistics for Particular Structures.....	82
4.20. ANOVA for Particular Structures.....	83
4.21. Scheffe Post-hoc Tests for Particular Structures.....	84
4.22. Correlations between Particular Structures and the Quality of L2 Writing.....	86
4.23. Descriptive Statistics for Lexical Richness.....	89
4.24. ANOVA for Lexical Richness.....	90
4.25. Scheffe Post-hoc Tests for Lexical Richness.....	91
4.26. Correlations between Lexical Richness and the Quality of L2 Writing.....	92

LIST OF ABBREVIATIONS

MLS: Mean length of Sentence

MLC: Mean Length of Clause

MLT: Mean Length of T-unit

C/S: Clauses per sentence

C/T: Clauses per T-unit

CT/T: Complex T-units per T-unit

DC/C: Dependent Clauses per Clause

DC/T: Dependent Clauses per T-unit

CP/C: Coordinate Phrases per Clause

CP/T: Coordinate Phrases per T-unit

T/S: T-units per Sentences

CN/C: Complex Nominals per Clause

CN/T: Complex Nominals per T-unit

VP/T: Verbal Phrases per T-unit

LR: Lexical Richness

L2: Second Language

SPSS: Statistics Package for Social Sciences

LFP: Lexical Frequency Profile

EFL: English as a Foreign Language

ESL: English as a Second Language

WECIL: Written English Corpus of Iranian Learners

CLAN: Computerized Language Analysis Program

TWE: Test of Written English

CHAPTER ONE:

INTRODUCTION

1.1. Background

As the world is getting ever more interdependent and increasingly globalized, the need to communicate through writing is growingly more felt. The advents of new technologies and interest in communicating via computers have attached a renewed importance to second language writing. To teachers' despair, writing is an anathema to most of EFL learners. The majority of researchers agree that "*students do not and cannot write well*" (p. 3) and EFL learners find it one of the arduous and demanding academic tasks (Amiran and Mann, 1982). Second language writing is overwhelmingly acknowledged as the most problematic skill in second language learning (Lee, 2005). Students are urged to produce well-written essays in order to be academically successful. Therefore, a question which merits considerable attention is: What issues are involved for producing a well-written essay? To answer this broad question, researchers have been trying to examine second language writing from different perspectives and dimensions. Even though there were some researchers like Angelova (1999) who identified a number of factors for the quality of L2 writing, it is not wise to confine the complex multi-componential phenomenon of second language writing to a superficial list of factors which may contribute to the quality of L2 writing.

It was not long time ago that Stephen Krashen stated that "*studies of second language writing are sadly lacking*" (1984, p. 41). Today there is a vast literature on second language writing. However, the contrasting findings of research studies, multifaceted nature of writing, and different

dimensions involved in second language writing persuade every researcher that second language writing is still under-researched. In spite of the fact that the past huge literature has been focused on different areas of second language writing, there are too many dimensions involved in each of these areas that make it so complicated to simply draw some conclusions from the confusingly accumulated findings. This is the reason why there are still some critical unremitting debates in the field of second language writing that need to be scrutinized closely.

In general, second language writing pundits consider three central dimensions for their research investigations (Cumming, 2001). Table 1.1 presents the main perspectives of second language writing scholarship:

Table 1.1.

Second Language Writing Dimensions & Perspectives

	Micro	Macro
Text	Syntax and Morphology Lexis	Cohesive Devices Text Structure
Composing	Searches for Words and Syntax Attention to Ideas and Language Concurrently	Planning Revising
Context	Individual Development Self-image or Identity	Participate in a Discourse Community Social Change

Adopted from Cumming 2001 (p.3)

The first dimension has to do with textual features of written texts. The second dimension is concerned with the composing processes that are involved in second language writing and the third dimension which has

recently started to gain a waxing popularity among second language researchers are about the sociocultural and sociopolitical contexts in which L2 texts are produced. When considered for investigation, each dimension can be looked at from a micro or macro point of view. Each perspective scrutinizes different incremental issues in second language writing.

1.1.1. Textual Features

Considering textual features being directly related to the product of writing, scholars have long been trying to analyze textual features from micro and macro perspectives. At the macro-level perspective, cohesive and organizing devices are considered for investigation (see for example, Allison, 1995; Jacobs, 1982; Grant and Ginther, 2000; Hyland & Milton, 1997; Reid, 1992, Reynolds, 1995). Many studies have investigated text types and structures such as argumentative writing (Connor & Farmer, 1990; Varghese & Abraham, 1998; Vedder, 1999), autobiographies (Henry, 1996) and narrative writing (Bardovi-Harlig, 1995). Within micro-level of discourse, however, some scholars argue for grammatical complexity and syntactic accuracy (see for example, Archibald, 1994; Bardovi-Harlig, 1995, 1997; Bardovi-Harlig & Bofman, 1989; Cumming and Mellow, 1996; Dickson, et. al., 1987; Harley & King, 1989; Ishikawa, 1995; Mellow and Cumming, 1994; Perkins, 1980; Reid, 1992; Sweedler-Brown, 1993; Weissberg, 2000) while others lay emphasis on lexical features (see for example, Engber, 1995; Grant & Ginther, 2000; Laufer & Nation, 1995, 1999; Reid, 1986; Sweedler-Brown, 1993; Cumming and Mellow, 1996).

1.1.2. Composing Processes

The second main dimension of second language writing is concerned with the composing processes. This dimension has also been studied from

two perspectives. The micro-level of perspective is about studying learners' thinking and decision-making episodes while writing a text. These studies have been trying to find some attributable salient writing behaviors among those who can write well. Some of them were focused on investigating learners' recurrent or fluent searches for proper words and phrases (see for example, Butler-Nalin, 1984; Chenoweth and Hayes, 2001; Cumming, 1990; Silva 1992; Uzawa, 1996). Some studies tried to examine learners' attention to ideas and linguistic forms at the same time while composing (see for example, Bell, 1995; Swain and Lapkin, 1995; Vignola, 1995; Whalen and Menard, 1995). These micro-level decision-making occurs in brief and temporary episodes. At the macro-level of perspective, studies have been concentrated on planning (see for example, Akyel, 1994; Roca de Larios, Murphy and Manchon, 1999; Zimmerman, 2000), revision (see for example, Hall, 1990; Urzua, 1987; Zamel, 1983) and finally editing (see for example, Walters and Wolf, 1996). In general, these studies concluded that more experienced, better L2 writers use these strategies more frequently compared with their less skillful peers.

1.1.3. Writing Contexts

The last dimension of second language writing is the most recent one. It has to do with the sociopolitical contexts in which people write. Like previous dimensions, it can be looked at from micro or macro perspectives. Research studies have been some case studies investigating the problems and difficulties a learner or a small number of learners have in a particular "social" context. The past studies also have investigated social contexts such as academic places (see for example, Angelova and Riazantseva, 1999; Casanave, 1995; Johns, 1992; Currie, 1993; Losey, 1997; Riazi, 1997;

Spack, 1997; Zamel, 1995), at home or public places (Cumming and Gill, 1991; Long, 1998), at work with certain vocational functions (Parks, 2000; Thatcher, 2000) and those who tried to get their works published in second language (Casanave, 1998; Gosden, 1996; Matsumoto, 1995).

1.2. Complexity:A Complex Component

This expansive bulk of research implies the importance and complexity of second language writing at the same time. In addition to this categorization of second language writing research, second language proficiency is also considered to be multi-faceted. There are three principal components recognized for second language proficiency namely accuracy, fluency and complexity (Skehan, 1998; Ellis, 2003, 2008; Ellis and Barkhuizen, 2005). These dimensions are most directly related to textual features. In other words, written products can be openly examined in terms of accuracy, fluency and complexity. As stated above, textual features can be researched from two perspectives. The micro-level is related to lexis and syntax while the macro-level is concerned with cohesive tools and structural matters. There has long been a heated debate on the position of grammar and vocabulary in language learning including second language writing. They can be analyzed in terms of accuracy, fluency and complexity. However, as implied by its term, complexity is the most complicated and ambiguous dimension. Ironically, this dimension has been unduly neglected and unjustifiably underestimated by some researchers and teachers (Skehan, 2001; Robinson, 2001). Consequently, it seems that this area of second language writing deserves a thorough exploration.

Complexity can be understood in two broad ways: cognitive complexity and linguistic complexity. Cognitive complexity is about how difficult and

complex the L2 system and its linguistic features are acquired and performed by L2 learners while linguistic complexity has to do with the L2 system and its features on its own (see for example, Dekeyser, 2008; Housen et. al., 2005; Williams & Evans, 1998). Grammatical and lexical features are among the most important properties of the L2 system (Ortega, 2003; Laufer& Nation 1995).

Grammatical complexity is generally concerned with the variety and sophistication of forms within a spoken or written text (Ortega, 2003). Given this vaguely-defined construct, it engulfs a myriad of characterizations as well as different methods for measurement. Lexical richness, on the other hand, is about how varied, sophisticated and rare words are used in a text (Laufer& Nation, 1995). Although there is a relative consensus on the broad definition of lexical richness (Engber 1995), the definition itself is wide open to interpretation. Complexity in structure and vocabulary of writing, therefore, seems problematic since there are still many issues within these two constructs that need to addressed methodically.

1.3. Grammatical Complexity

Grammatical complexity has been proposed as an important scale to measure how complex and sophisticated production units are in a piece of writing (Foster & Skehan 1996, Ortega 2003, Wolfe-Quintero et al. 1998). Regrettably, the findings of the past studies are not in any way conclusive and even sometimes contradictory in this area. On the one hand, there are researchers who have found significant associations between language proficiency or writing proficiency and the construct of grammatical complexity (Cooper, 1976; Elova, et al., 1980; Hirano, 1991; Kameen, 1979; Kawata, 1992; Monroe, 1975) while others, on the other hand, did not find

any association of significance at all (Bardovi-Harlig, 1992; Bardovi-Harlig&Bofman, 1989; Perkins, 1980; Perkins & Leahy, 1980; Sharma, 1980; Tapia, 1993; Vann, 1979). There are also some studies that have reported mixed results about this relationship (Flahive& Snow, 1980; Homburg, 1984; Ishikawa, 1995). Perhaps these sheer inconsistencies can be explained on the account of methodological differences, different interpretations of grammatical complexity and different situations and contexts under which studies have been conducted. Since grammatical complexity has not been clearly defined in the literature, each researcher has come up with their own reading of this construct leading to a multitude of measures.

Conceptually, the main measures of grammatical complexity can fall into five types (Lu, 2010): Mean length of production unit, sentence complexity, subordination, coordination and particular structures. For the analysis of these five types of grammatical complexity, fourteen measures can be employed i.e. mean length of sentence, mean length of T-unit and mean length of clause (for mean length of production unit); clauses per sentence (for sentence complexity); T-unit complexity, complex T-unit, dependent clauses per clauses and dependent clauses per T-unit (for subordination); coordinate phrases per clause, coordinate phrases per T-unit and sentence coordination (for coordination); and complex nominals per clause, complex nominals per T-unit and verb phrases per T-unit (for particular structures). These measures are summarized in Table 1.2:

Table 1.2.

Types and Measures of Grammatical Complexity

No.	Measure	Code	Type
1	Mean Length of Sentence	MLS	Mean Length of Production Unit
2	Mean Length of Clause	MLC	Mean Length of Production Unit
3	Mean Length of T-unit	MLT	Mean Length of Production Unit
4	Clauses per Sentence	C/S	Sentence Complexity
5	T-unit Complexity	C/T	Subordination
6	Complex T-unit	CT/T	Subordination
7	Dependent Clauses per Clause	DC/C	Subordination
8	Dependent Clause per T-unit	DC/T	Subordination
9	Complex Phrases per Clause	CP/C	Coordination
10	Complex Phrases per Clause	CP/T	Coordination
11	Sentence Coordination	T/S	Coordination
12	Complex Nominals per Clause	CN/C	Particular Structure
13	Complex Nominals per T-unit	CN/T	Particular Structure
14	Verb Phrases per T-unit	VP/T	Particular Structure

It is worth noting that these measures are not by any means exhaustive. They are only those measures that have been pegged as important measures of language development in the literature. There are over thirty measures of grammatical complexity suggested in the past that we will discuss them in situ.

1.4. Lexical Richness

Lexical richness is concerned with how varied, sophisticated and

advanced words are used in writing or speaking. But lexical richness is an umbrella term which involves different lexical aspects of complexity. Even though some researchers have acknowledged the significance of lexical richness in second language writing performance (McClure, 1991; Engber, 1995; Laufer, 1994; Laufer & Nation, 1995; Coniam, 1999; Goodfellow et. al., 2002; Li, 1997), there are many researchers who did not find any relationship between lexical richness and proficiency (Cumming & Mellow, 1996; Arthur, 1979; Hyltenstam, 1988; Nihalani, 1981; Linnarud, 1986). There might be many reasons behind these contradictory results, but one reason is for sure about the way lexical richness has been approached in each study. Given the definition of lexical richness, a good analysis of lexical richness must analyze all of its aspects including lexical density, lexical diversity and lexical sophistication. Lexical density has to do with the number of content words in a text. The more content words, the denser and richer the text will be. It was first proposed by Ure (1971) as a measure of lexical richness. This aspect of lexical richness can be simply calculated through the ratio of the number of lexical words to the total number of words in a piece of text. The second aspect is about lexical diversity. Lexical diversity (sometimes called lexical variety) has to do with the range or size of a learner's vocabulary as manifested in his/her writing and speaking performances. It can be calculated through the ratio of type and token. The last aspect of lexical richness is termed as lexical sophistication (sometimes called lexical rareness) which has to do with "the proportion of relatively unusual or advanced words in the learner's text" (Read, 2000, P. 203). These measures of lexical richness are summarized in Table 1.3.

Table 1.3.

Measures of Lexical Richness