UNIVERSITY OF CENTRAL OKLAHOMA
Edmond, Oklahoma
Dr. Joe C. Jackson College of Graduate Studies

The Use of And-Coordination in terms of Its Syntactic (A)symmetry in Argumentative Essays: A Corpus-Based Study of Three University Learner Groups in MICUSP and NUCLE

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements

for the degree of

MASTER OF ARTS IN ENGLISH

By

NhuQuynh Luu Nguyen

Edmond, Oklahoma

2013

The Use of And-Coordination in terms of Its Syntactic (A)symmetry in Argumentative Essays: A Corpus-Based Study of Three University Learner Groups in MICUSP and NUCLE

A THESIS

APPROVED FOR THE DEPARTMENT OF ENGLISH

June 2013

By _____
Dr. John Hitz, Chairperson

_____
Dr. Mary Spelman, Member

_____
Dr. Matthew Hollrah, Member

ACKNOWLEDGEMENTS

Corpus linguistics captivated me in the first semester I joined the TESL's program, and ever since, I have envisioned utilizing corpora for teaching English grammar and writing to speakers of other languages. Toward the goal I wrote this thesis, during which I have interacted with and received tremendous support from professionals, friends, and my family. With great gratitude, I realized I could not have finished the thesis without them, and what I received from them defies any temporal limitation as it paves the way for my participation in corpus research, and lies within me like a heartbeat.

I would like to thank my committee chair and mentor, Dr. John Hitz, for marvelously continuous encouragement and perceptive guidance, with which I grew my ideas into a feasible research, improved the drafts, and persisted until its completion; my committee advisor, Dr. Matthew Hollrah, for acute feedback by which I was more aware of the writing process to present my topic with clarity; my academic and committee advisor, Dr. Spelman, for essential suggestions since the very beginning when I opted for a corpus research for my thesis track; Dr. Laura Dumin for comments on an embryo of my research idea, by which I gained more confident in my choice of the topic and knew which aspect is interesting to explore more; Dr. Hilary Nesi (Coventry University) for insights about the influence of genres in writing and the concepts "hypotaxis/parataxis"; my dear cousin, Dr. Ngan Luu Thuy Nguyen (University of Tokyo), for introducing the NUCLE corpus, which I found more suitable than others for the research purpose; and the NUCLE corpus' author, Dr. Daniel Dahlmeier, for general information about the writers in the NUCLE corpus.

This thesis is dedicated to my beloved parents, Mai Kim Luu and Dung Thanh Nguyen, whose unconditional love keeps me move forward; my elder brother, Doan Bao

Luu Nguyen, and my sister-in-law, Thy Nguyen Lam Xuan, for supporting me all the way and preparing me what to expect in thesis writing; Dr. David Macey, who helped me earn sponsorship for my trip to a conference in corpus linguistics, for which I had opportunities to meet with experts in the field; Ms. Alison Hedge Coke for delicacy and care; my wonderful friends: Alisha Costanzo Chambers, Matt Garley (Freiburg Institute for Advanced Studies), Eunlee Ryu, Ms. Darla Sherman, and my co-workers at Tutoring Central.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT OF THESIS

AUTHOR: NhuQuynh Luu Nguyen

TITLE: The Use of And-Coordination in terms of Its Syntactic (A)symmetry in Argumentative Essays: A Corpus-Based Study of Three University Learner Groups in MICUSP and NUCLE

DIRECTOR OF THESIS: Dr. John Hitz

PAGES: 131

Studies found EL learners overuse *and* as an additive connector at the sentence-initial position (Bolton, Hung, & Nelson, 2002), and they underuse *and* as a coordinator (Leung, 2005). Generally, the use of the *and*-coordinator has often been overlooked in corpus research and in English teaching because of its seemingly simplicity. To test previous findings about the *and*-coordinator and to examine the influence of English proficiency on the use of *and* in academic writing, three learner corpora—MICUSP-NNS (advanced level), MICUSP-NS (advanced level), and NUCLE-NNS (upper-intermediate) were compared, with regard to the use of (a)symmetric structures of the *and*-coordination. Each corpus contains 31 argumentative essays written by 31 university students.

One-way ANOVA and Kruskal-Wallis tests indicated more evidence is needed to determine whether learners at advanced and upper-intermediate levels use total *and*-coordinators differently. More evidence is also needed to decide if the three groups use the *and*-syntactically symmetric coordination differently. Students at the upper-intermediate level, however, use *and*-asymmetric structures significantly less than those at the advanced level. A subsequent qualitative analysis reveals their use of *and* shows more repetition and redundancy. Possible reasons for why they employ asymmetric *and*-coordination much less frequently includes the lack of knowledge of certain grammatical

structures, the insufficient knowledge of the writing topic, the impact of instructions, and language transfer.

An implication of the study is *and*-coordination should be taught at all levels of proficiency and should be integrated into the teaching of other grammatical structures. When students' use of *and* shows unskillful repetition and redundancy, teachers might need to help students acquire knowledge of other structures and the academic topic they are writing about. Future studies can examine how differently learners at various proficiency levels use *and* in connecting different types of phrases and clause and what the relationship between parallelism and syntactic asymmetry in the *and*-coordination is.

The Use of And-Coordination in terms of Its Syntactic (A)symmetry in Argumentative Essays: A Corpus-Based Study of Three University Learner Groups in MICUSP and NUCLE

Chapter 1: Introduction

1.1. Purpose of the thesis

The aim of this thesis is to examine in what way English proficiency influences the use of *and*-coordination in academic argumentative writing by college students. The thesis attempts to analyze whether students, including native speakers at an advanced level, non-native speakers at an advanced level, and non-native speakers at an upper-intermediate level, demonstrate differences in use of the syntactically symmetric and asymmetric (henceforth (a)symmetric) structure of *and*-coordinator. The ultimate goal of the findings is to identify the needs of ESL college learners in academic argumentative writing.

1.2. Importance of the thesis

Throughout, the thesis underlines issues that emerge quite often but have not yet been settled or proportionately highlighted in corpus research. Its theoretical, empirical, and methodological contributions also purport to strengthen the interrelation between Corpus Linguistics studies and other fields, including Second Language Acquisition and Teaching English to Speakers of Other Languages (TESOL).

Theoretically, the thesis offers a synthesis perspective about the *and*-coordination to elucidate what it means by the symmetric nature of *and* and the syntactic (a)symmetry, and how these concepts are manifested in learner language. With respect to second language acquisition, the findings clarify the role of second language (L2) proficiency in learners' use of *and*-coordinators in academic writing, and might indicate an interaction between L2 proficiency and forward transfer. An implication of the results to TESOL is

how English teachers should teach coordination in academic writing to learners of different proficiency levels.

Empirically, the inferential statistics, including parametric and non-parametric tests to cover the possibility of a non-normal data, substantiate previous findings that *and*-coordination is underused (see Leung's study, 2005), and cast doubt on the common belief that learners' overuse *and.*

Regarding methodologies, terms that are often mistaken or ignored in many corpus studies, such as *genre* and *text type* or *corpus-based* and *corpus-driven*, are clarified. The study also promotes a common descriptor of learners' proficiency for the purpose of comparison across corpora.

1.3. Outline of the thesis

The thesis begins with an introduction that evaluates its importance in the field of corpus linguistics, grammar, second language acquisition, and teaching English as a second language. It recapitulates the main points each chapter discusses and the rationale to arrange the content as it is. The major research questions and the results are presented in brief, and relevant terms, including text, cohesion, genre, and corpora are defined in detail to lay ground for more in-depth discussion in subsequent chapters.

Chapter 2 presents the current trend of research on cohesive devices in general and the conjunction *and* in particular. Much research has been done to clarify what cohesive devices English learners from different first language (L1) backgrounds overuse, underuse, and misuse, and wherein they use them. The cohesive devices receiving enormous attention are mostly logical connectors. The concept of *overuse, underuse,* and *misuse* is upheld depending on what researchers adopt as the reference for

comparison with learners' production and whether inferential statistics detect any significant difference between learners' and native speakers' usage patterns. Typically, the reference is native speakers' spoken or written texts, but some argue only experts' should be considered the norm for comparison.

The chapter then zeros in on research of *and*-conjunctions in writing. Most studies found *and* as a logical connector is overused; a few concluded *and*-conjunction is underused. A further investigation that goes beyond the difference in frequencies of *and* involves learners' acquisition of *and.* Within it, two key issues are discussed: (language) *forward* transfer (i.e., how L1 influences L2 usage) and relation between hypotaxis and parataxis. Another issue, namely, how frequencies of *and* are regulated by genres, receives considerable attention.

This chapter is a key stage whereby the current study is characterized. First, the author's analysis includes a comparison of frequencies of *and* between non-native speakers' and native speakers' writing, but the concept *overuse* or *underuse* is not applicable. Second, since the main criterion for the comparison is *proficiency level*, other factors that might influence the results, including genres and language transfer, are either controlled or analyzed in detail. Third, the author addresses the controversy around methodologies in corpus research and adopts a stance. The study, thus, is a combination of qualitative, quantitative, corpus-based, and top-down approaches.

In chapter 3, the author presents how Transformational Grammar and Functional Grammar delineate the concept "*and*-coordinator," and synthesizes the two perspectives to classify *and*-coordinators into two groups: syntactical asymmetry versus syntactical symmetry. The researcher utilizes the description of the two groups to categorize

instances of *and* in learners' writing. The chapter ends with a preliminary error analysis of the NUCLE corpus to test the classification and to establish a general idea for formulating the hypotheses. A brief justification of the research at the end explains why answers to the research question are meaningful.

The final section in this chapter presents the major research question, namely, wherein learners of various proficiency levels use *and*-coordinator with regard to its frequency and syntactic features. The researcher hypothesizes that there is at least one group using total *and,* as well as *and*-syntactical (a)symmetry, significantly different from the other two groups. In this research, the independent variable is the English proficiency level, and the dependant variable is the frequency of *and*-coordinators.

A quantitative research design is set up and described in chapter 4 to address the hypotheses. According to the design, the researcher investigates frequencies of *and*-coordinator in argumentative writing across three learner groups: MICUSP-NS, MICUSP-NNS, and NUCLE-NNS. The chapter continues with the discussion of key issues related to the samples, i.e., what constitutes an argumentative essay, how proficiency levels of the three groups are determined, what *English proficiency* in *argumentative academic essays* is, what the sampling procedure of this research is, what we know about the topics and disciplines of the essays, and to what extent the three corpora are compatible for comparison.

In chapter 5, descriptive statistics and inferential statistics reveal differences in terms of frequencies of *and*-coordinators, but only in the category of *and*-asymmetric structures, statistically significant results were found: the NUCLE-NNS group use *and*-asymmetry significantly less than the other advanced groups.

The concluding chapter 6 reinstates the findings, formulates explanations for each finding, confirms the contribution of the thesis to research, considers possible limitations, and presents suggestions for future research and the teaching of *and*-coordination.

1.4. Research questions and results

The major research question is as follows: *in what ways university learners at upper-intermediate to advanced levels of English proficiency, represented by the three groups (MICUSP-NS, MICUSP-NNS, and NUCLE-NNS), use and-coordinators in argumentative academic essays, as long as frequencies of and-coordinator and its syntactically (a)symmetric features are the main concern.*

The results reveal the non-native advanced learners—the MICUSP-NNS group—used the total *and*, as well as *and*-syntactic (a)symmetric structures, slightly more (but statistically non-significant) than the other two groups. The non-native upper-intermediate learners—the NUCLE-NNS group—used the total *and*, as well as the *and*-symmetric coordination, slightly less (but statistically non-significant) than the other two advanced groups. The NUCLE-NNS group, however, used *and-a*symmetric structures much less than the MICUSP-NNS and the MICUSP-NS groups.

1.5. Definition of key terms

The purpose of this section is to present key concepts that have received the spotlight in the studies of cohesive devices in written texts by learners (native and non-native speakers of English), as well as those relevant to my present study.

1.5.1. Text, cohesion, and cohesive devices

Consider the following passage:

> Hegel wrote, "When we turn to survey the past, the first thing we see is nothing but ruins." As I contemplate the development of the coast, looking at old photographs of once new building – the pride of a growing city – I see beneath them, as if a palimpsest, the destruction wrought by Katrina. (Trethewey, 2010, p. 51)

The excerpt can be perceived as an English text that has a sense of "flow" from the first sentence to the next and within each one: Hegel's statement reflects the reality the writer was experiencing; she realized after Katrina, what was left of once new building—a sign of coastal development—is nothing but old photographs and ruins. The sense of "flow" that prompts readers to perceive the two separate sentences as well as the clauses within a sentence as a meaningful, unified whole is what Williams called "cohesion" (2010, p. 68).

Cohesion is not a new linguistic phenomenon. According to Xi (2010), the antecedent of cohesion study in English was related to syntax and parallelism in poetic texts. Later research attempted to describe features of cohesion and its potential application in stylistics. Not until Halliday and Hasan (1976) published *Cohesion in English* did cohesion become a theory whose concept and classification of cohesive devices have been served as a framework for research on texts as well as discourse. Text, as the authors defined, is "any passage, spoken [(verbal record of communicative events)] or written, of whatever length, that does form a unified whole" (1976, p.1), and cohesion is a factor that transfers "texture" quality to a string of sentences so that they are not randomly concatenate.

In interpreting Halliday and Hasan, Butler (2003) emphasized that cohesion is actually seen as a semantic relation whereby the comprehension of one element in the discourse is dependent on that of another (p. 336). Halliday called the relationship between the two units a "cohesive tie," and pointed out its important aspect as "being non-structural," meaning that cohesive ties are not manifested through grammatical structures at sentential or clausal level but through "elements of any extent, both smaller and larger than clauses, from single words to lengthy passages of text, and that may hold across gaps of any extent…" (as cited in Butler, 2003, p. 336). Cohesion, he concluded, is rather a process, not a product, actualized by cohesive devices to achieve texture of text. He suggested two types of cohesion: grammatical cohesion, actualized by reference, substitution, ellipsis, and conjunction; and lexical cohesion, including reiteration, synonym, hyponymy, and collocation.

Butler's discussion on Halliday and Hassan's theory provides an important implication for the current research, that cohesion and cohesive devices are related but separate. In written texts, the number of cohesive devices may not correlate with the amount of text cohesion, but instead signals the writer's awareness of the actualization (*actualization* is my term to indicate the process of selecting a cohesive device to signify a cohesive relationship). The current study, although does not penetrate into the writer's psychological process, offers an attempt to explore further the relationship between cohesion, cohesive devices, and the writer's awareness of the actualization through the use of coordinator *and*. In this study, four elements come into play: the writer's English proficiency, academic writing, *and*-(a)symmetric coordination, and frequency of *and* conjoining (a)symmetric structures. The preliminary assumptions built in the study are

that *and*-asymmetric coordination may involve more complicated structures than *and*-symmetric coordination, and that if writers perceive complicated structures as difficult to produce or comprehend, then they may use them less as *and*-conjuncts. How these elements interrelate and how the assumptions are checked will evolve throughout to the end of the research.

The research is premised on Halliday's framework and Celce-Murcia's classification of cohesive conjunctions (1999) (Table 1).

Table 1

Type of cohesive conjunctions

Type of cohesive conjunctions	Conjunctions	Cohesive ties
Coordination (alternatively referred to as Coordinating Conjunction)	And	Plus
	Or	One or the other of two alternatives is true
	Nor	Conjoins two negative sentences, both of which are true
	So	Therefore
	But	Show contrast
	Yet	But at the same time
	For	Because
Logical connectors: Simple Adverbial Subordinators (alternatively referred to as Subordinating Conjunctions)	After, as long as, as soon as, before, since, when, whenever, until	Time
	Where, wherever	Location
	As, in that	Manner
	So that, in order that	Purpose
	Since, because, as, inasmuch as, now that	Reason
	While, as	Simultaneous
	If, even if, as long as, in case, provided that	Conditional
	Although, even though, though, while	Concessive
Logical connectors: Conjunctive Adverbial	In addition, moreover, furthermore, besides, also	Emphatic additive

(alternatively referred to as Adverbial Conjunction)	That is, in other words, for instance	Appositional additive
	Likewise, similarly	Comparative additive
	However, nevertheless, despite this, in contrast	Proper adversative
	In fact, actually, however, on the other hand, at the same time	Contrastive adversative
	Instead, rather, on the contrary, at least	Correction adversative
	In any case, anyhow, at any rate	Dismissal adversative
	Therefore, consequently, for that reason, thus	General causal
	Then, in that case, otherwise	Causal conditional
	Then, next, first, second, last finally, up to now, to sum up	Sequential

1.5.2. Relationship between cohesion and coherence

With Butler's summary of Halliday's dynamic perception on cohesion from 1976 to 1994, a discourse between Halliday and other critics such as Widdowson and Carrell can be established. Cohesion by Halliday is not confined to grammatical structures, but instead, to the semantic ties within a text. For some texts that represent discourse in communication, however, cohesive ties seem to be absent. Consider the following conversation:

A: Could you tell me the time?

B: Sorry, I forgot my watch.

Clearly, B's excuse does not constitute a cohesive tie with A's question because A was not interested in whether B forgot his watch or not. The whole conversation, however, makes sense in that A was making a request, and B's statement was an explicit reason for an implied reply to A's request. This is what Widdowson (1978) calls "illocutionary acts" (acts that are performed by utterances), and asserts that the text can be coherent without an overt cohesive tie.

Different from spoken discourse, writers may need to establish overt cohesive ties to enhance readers' comprehension of the text. They do so through maintaining logical connections between their ideas, regardless whether those connections are signaled by cohesive devices or not.

Carrell (1982) advanced on Halliday's idea about cohesion that a text can have cohesion but appear incoherent to readers (as cited in Butler, 2003). Odlin (1989) also noticed a possible negative transfer in grasping the coherence of a text is for people coming from different cultures and technical backgrounds, or for those who are not familiar with the patterns—the presentations of information. In fact, the audience comprehends a text based on not only the surface linguistic features that convey meaning but also his or her background and cultural knowledge of the topic. Negative transfer in coherence, however, would not be an issue for the present research, its sample being comprised of texts written by upper-level college learners and beyond.

The manifestation of cohesion in texts, in many circumstances, is also regulated by genres. It is therefore important to understand what *genre* is, especially in the context of corpus linguistic research.

1.6. Genres

1.6.1. Definition

The word "genre" was first used for literary forms, such as novels and poems, until Halliday developed Systemic Functional Linguistics (SFL) and included in its definition the intimate connection between culture, context and language. "Genre," according to Halliday, begins with "contexts of situation" which are not unique but often reoccur as "situation type, set of typified semiotic relationship," establishing a scenario of

people, actions, and events that prompts a participant in that scenario to interpret their meaning in a specific way (1978, p. 100). Adhering to the norms overtime, the participant solidifies the rules for communication—a typical way of linguistic interaction with the others—being what Halliday referred to as "specify the semantic configurations that the speaker will typically fashion" (p. 100).

To articulate meaning in a context of situation, human beings use language, which in turn helps them socialize and perform meaningful actions. Through interaction, participants form the "social semiotic"—a "network of meanings" embedded in any culture. This semiotic network, as Halliday explained, is maintained by its discourse-semantic systems to represent a culture's "meaning potential" (p. 113).

Halliday (1978) ascribed this "clustering of semantic features according to situation types" the name "register" (p. 68). Register connects a situation type with particular linguistic patterns to describe the event (the "field"), the relationship between people (the "tenor"), and the role of language (the "mode"). A scientific register, for example, encompasses stylistic choices, interactional patterns, and communicative means commonly used in scientific contexts (Bawarshi, 2010, p. 31).

Halliday's theory of genre, however, was not flexible enough to be immune to criticism: it was questioned by other research findings that one genre can have a wide variation of linguistic patterns, or different genres can share the same linguistic features. Swales (1990) introduced the term "discourse community" and defined genre as the pragmatic knowledge shared by discourse community members of communicative events with common communicative purposes (as cited in Bawarshi, 2010, p. 44).

Martin (1997, p. 37) expounded on Halliday's genre theory by emphasizing the connection of culture with situational types and language. A genre is how language is used in a specific context of situation that fits in within a cultural expectation, or, as Martin concluded, genre contextualizes register, which encompasses field, tenor, and mode, and register in turn contextualizes language.

1.6.2. Mode of discourse

As a related term to genre, modes of discourse—the surface text types that Longacre (1996) called genre—are Narrative, Expository, Behavioral, and Procedural. Modes of discourse can cover a wide range of genre. According to Diller (2001):

> …the Narrative mode overarches genres such as fairy tales, novels, short stories, newspaper reporting; the Procedural mode includes such genres as food recipes, how-to books, etc.; the Behavioral mode includes essays and scientific articles; and the Expository mode includes sermons, pep-talks, speeches, etc." (p. 13)

1.7. Corpus

1.7.1. Definition

Corpus is "a collection of texts—written, transcribed speech, or both—that is stored in the electronic form and analyzed with the help of computer software programs" (Conrad, 2005, p. 393). This definition implies what potential a corpus has in giving an insight to the authentic language, specifically, word frequencies and co-occurrences, derived from users of that language for communicative purposes. Some corpora also consist a variety of genres mostly classified by hand.

As Conrad mentions, corpora in the '70s consisting of 1 million words were considered large, but today, technological development allows researchers to build those

that are hundreds of times larger than before, such as the Corpus of Contemporary American English with 415 million words (p. 394). Although large corpora may represent the language sufficiently, she notices, smaller one can also be representative.

1.7.2. Genre in corpora: A revisit to the definition of "genre"

The process of compiling and analyzing corpora requires an understanding of *genre* to classify texts, but the notion of genre is often used interchangeably with other related terms, including text type, register, domain, style, sublanguage, message form, and so forth (Lee, 2001). Thus, a clarification of *genre* in comparison with other most common terms, and a summary of criteria for classifying texts are necessary in the present research to evaluate the compatibility of the corpora under examination.

Genre classification is dependent on non-linguistic factors, what Lee called external elements, e.g., the intended audience, purpose, and activity type. Biber's and Pattridge's text type, in contrast, colligates internal, linguistic characteristics of texts themselves, i.e., surface-level lexicogrammatical or syntactic features as in Biber's definition, or internal discourse/rhetorical structure type as the determinant according to Pattridge's definition (as cited in Lee, 2001). A text type fitting Biber's description could be "novels" or "biographies," and Pattridge's could be "argumentation" or "narration."

From another perspective, genre and register, Lee suggested, are "two different points of view covering the same ground" (p.46). Register is considered when people view text as language while genre comes into play when people perceive text as a member of a category or a cultural artifact. That is why *formal register* makes sense, but formal genre is not. This perception resonates with the aforementioned Systemic Functional Linguistics' viewpoint, conceptualizing genre as the "organization of culture

and social purposes around language" and register the "organization of situation of immediate context" (p.46).

Given a fuzzy set of boundaries to distinguish genre from other terms, the Prototype Approach attempts to characterize genres by their cognitive statuses into basic-level, sub-genre, and super-genre. Genres that fall into the basic-level status can be distinguished more easily in terms of seven attributes (which according to Lee are more in favor of written genre):

> …*domain* (e.g., art, science, religion, government), *medium* (e.g., spoken, written, electronic), *content* (topics, themes), *form* (e.g., generic superstructures, à la van Dijk (1985), or other text structural patterns), *function* (e.g., informative, persuasive, instructive), *type* (the rhetorical categories of "narrative," "argumentation," "description," and "exposition") and *language* (linguistic characteristics: register/style). (p.49)

Based on Lee's synthesis, I propose the following criteria whereby the corpora in this research will be compared (Table 2).

Table 2

Criteria for text classification in the corpora

Criteria	Example
Medium	Spoken, written, electronic
Function	Informative , persuasive, instructive
Text Type	Narrative, argumentation, description, exposition
Audience level	Low, medium, high
Genre	Student's essay, news report, letter
Discipline family[a]	Arts & Humanities (AH), Life Science (LS), Physical Sciences (PS), Social Sciences (SS)

Note. [a]This category is mentioned in Nesi's study (p.8, 2012). *Discipline family* is chosen instead of *domain* because it implies the essays being compared are writings across disciplines.

1.7.3. Developmental language corpora

There are many types of corpora, such as diachronic (presenting language throughout a period of time) versus synchronic (presenting a temporal "snapshot" of language), annotated (linguistic analysis encoded in the data) versus unannotated (no predetermined linguistic encoding), and general (aiming to represent the language as a whole) versus specialized (aiming to represent a genre or other segments of language) (Lee, 2010; McEnery, 2012). The present research focuses on developmental written language corpora, providing a general background of the category to which MICUPS and NUCLE, the two corpora under this analysis, belong.

Developmental language corpora, or *learner corpora*, are collection of various data from non-expert language produced by native and non-native speakers (Lee, 2010). Its alternative name, *learner corpora*, often denotes a sense of *foreign learners* rather than all learners, and hence will be used only when referring to corpora comprised exclusively of ELL's language output. Researchers often create their own corpora to fit their research design, but there are a few ready-made, large-scale developmental corpora that were considered candidates for this research.

The Louvain Corpus of Native English Essays (LOCNESS; 324,000 words) contains argumentative essays by American and British students, post-secondary to university levels (Lee, 2010). The British students' essays were collected in 1991, and the American students' essays in 1995 (Gilquin, 2011). LOCNESS is often used in comparison to the ICLE Corpus of non-native speakers' writing because they share the same genre and topics (Lee, 2010).

The ICLE corpus (currently about 3.7 million words) is a combination of sub-corpora created by different partner universities since 1990. The sub-corpora follow the

same format, and contain argumentative essays written by higher intermediate to advanced non-native college students from fourteen L1 backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish) (Gilquin, 2011). Essays in both the ICLE corpus and LOCNESS are mostly for composition classes (Granger, 2003).

With regard to writing for academic disciplines, not for composition classes, there are two main corpora, the British Academic Written English (BAWE) and the Michigan Corpus of Upper-Level Student Papers (MICUSP) (Lee, 2010). The BAWE corpus consists of 6.5 million words from 2858 upper-level essays written in thirty five disciplines in three universities: Universities of Warwick, Reading and Oxford Brookes (BAWE's manual). Its American counterpart, the MICUSP corpus, is smaller, with a total of 2.6 million words from 830 A-graded papers, written by the University of Michigan students, across sixteen disciplines. Both corpora include the writing of non-native and native speakers of English.

Despite being a counterpart of the MICUSP corpus, BAWE is scarcely used as comparable to MICUSP. One reason could be the classification of texts based on genres and other linguistic factors in the two corpora is varied; and, therefore, the texts in BAWE may unfold different writing moves and stages as compared to those in MICUSP of the same genre. My resolve is to compare two sub-corpora derived from the same corpus to ensure that the selected texts are more alike in terms of genre classification.

Because the study's focus is about students' academic writing, MICUSP seems to be the best fit, and for this study, texts in MICUSP were separated into two sub-corpora, one including native speakers' and the other non-native speakers of English. A third

learners' corpus, the NUCLE corpus, were added to create three levels of language proficiency—also the independent variable of the study. Further discussion on the NUCLE corpus will be presented in Chapter 4.

Chapter 2: Background of the research

The purpose of this background section is to position, or more precisely, define the boundary of my present research in the context of other studies of cohesive devices in writing texts by learners (native and non-native speakers of English)

Previous research on cohesive devices in learners' texts with a focus on advanced learners follows three main directions: in what cases learners overuse, underuse or misuse them; how the use of cohesive devices varies across genres; and what genre analysis of texts reveals about the cohesive relationships within texts. The third direction, which involves using Rhetorical Structure Theory to examine complex cohesive relationships, whether they are actualized (signaled) by cohesive devices or not, is irrelevant and hence outside the scope of this study whose main concern is restricted to one cohesive device only, the *and*-coordinator.

2.1. Overuse, underuse, or misuse of cohesive devices

The proliferation of research on misuse, overuse and underuse connectives reflects the need to recognize the areas that learners struggle with in constructing text cohesion. While overuse and underuse of connectives may cost extra cognitive processing time, hence increasing reading time for the text, misuse may hinder text comprehension (Crewe, 1990; Ben-Anath, 2005). Various studies cited by Chen (2006) as well as by Alarcon and Morales (2011) identified types of connectors that English language learners overuse, underuse, or misuse, by comparing their essays with those written by native speakers or native experts of English. The first languages of those learners of interest includes Chinese (Crewe, 1990; Bolton, Nelson, & Hung, 2002; Braine and Liu, 2005; Field and Oi, 1992; Field and Yip, 1992; Meisuo, 2000; Milton

and Tsang, 1993; Lei, 2012; Yeung, 2009), Malay (Johnson, 1992), French (Granger and Tyson, 1996), Sweddish (Altenberg and Tapper, 1998), British (Bolton et al., 2002), and Taiwanese (Chen, 2006).

The aforementioned studies on Chinese learners' essays in English showed that Chinese EFL students at various academic levels, from high school to undergraduate and graduate, often overused or misused linking adverbials. They most frequently overused sequential (firstly, secondly, lastly), emphatic additive (besides, moreover, furthermore, also), and causal (therefore, consequently) conjunctions. Among causal linkers, however, *thus* and *as a result* are rarely used. Unlike native speakers, L2 learners usually positioned connectors at the sentence initial. Misuse happened most often in the adversative category, when learners employed *on the contrary* for *however* or *on the other hand, on the other hand* and *however* for additive relations, and *in fact* without any implied contrastive nuance.

Compared to Chinese EFL learners, students from other L1 backgrounds have similar issues but with slightly different patterns. French EFL learners tended to underuse *however, therefore* and *thus* while overusing *moreover, for instance* and *on the contrary* (Granger and Tyson, 1996). Advanced Swedish EFL learners often avoided using conjunctions, e.g., *hence, therefore, thus,* and *however,* in their writing (Altenberg and Tapper, 1998). British students overuse *however, so, therefore, thus* and *furthermore* (Bolton et al., 2002). Advanced Taiwanese EFL learners misused additive connectors in formal register—connectors such as "besides," "plus," "what's more," and "actually" that should not be used in formal writing (Chen, 2006). Hinkel (2001) analyzed English essays written by Korean, Japanese, Indonesian, and Arabic, and concluded that the essays

contained significantly more sentence transition devices than NSs', but the higher use did not guarantee more coherence in discourse because it was often the case that transition devices in NNS's texts forcefully connected somewhat disjointed ideas.

Regarding the use of connective *and* in EFL essays, research comes to a unified conclusion that *and* is among the top frequently-used connectors, and the *additive* category, which includes *and*, is the most common one. Alarcon and Morales (2011) found that addition positive (e.g., and) comprised 20.86% of the total conjunctions, following addition negative (e.g., but)—the most popular that accounted for 23.62%. They also noticed that additives use was limited to *also* and *and*; the other alternatives such as *moreover, furthermore, in addition* were much less frequent. Their study's result resembled a previous qualitative analysis by Meisuo (2000) in which she found a tendency of overusing additives and temporal devices by Chinese learners, and Bolton et al.'s quantitative research (2002), which demonstrated that both Chinese and British university learners overused the *and*-connector, but the British used less *and* than the Chinese group. It should be noticed that the *and*-connector in Bolton et al.'s study only included the *and*-additive conjunction at sentence-initial positions.

While native English speakers' writings were de facto the reference of NNS's in determining which cohesive devices NNS overuse, misuse, or underuse, the above mentioned studies differed in the academic level of native writers. Bolton et al. (2002) preferred published academic texts (e.g., peer-reviewed journals) to students' writings as the reference, arguing professional texts provide the best cohesion model.

Among the studies in this background section, some did not include NS's texts as a reference because they were more concerned how the number of cohesive ties

influences the quality of the essay, hence the unnecessary of using NS' texts. Applying Pearson correlation and analysis of variance on EFL Chinese speakers' texts, Meisuo (2000) determined there were neither a significant relationship between cohesive ties and the essays' quality nor a significant difference in the quantity of cohesive ties between the highly-rated and the poorly-rated compositions. In contrast, Martinez (2004) inferred from her sample of EFL Spanish speakers' expository essays that compositions' grades were correlated with the number of cohesive devices (or discourse markers), and contrastive and elaborative discourse markers appeared more often in highly-rated texts than in low-rated ones. Braine and Liu (2005) came to the same conclusion for Chinese students' essays, but included the lexical category as a type of cohesive devices, while Alarcon and Morales (2011) excluded it and found the opposite result.

In general, the research on misuse, overuse and underuse of connectives incline to be inclusive; each study attempted to cover as many types of cohesive devices as possible and focused on their disparity in frequency or ratio of frequency use. Some used the qualitative method as a follow-up of the quantitative analysis to explore features of individual connector's usage in texts, for example, their positions—whether they are sentence-initial or sentence-final—as in Lei's study (2006). Such an all-encompassing and follow-up approach is beneficial in revealing the interrelation among cohesive linkers, but diverts attention from analyzing individual linkers thoroughly to understand the syntactic and semantic differences across groups of learners' writings. The researchers in those studies on learners' use of *and* found the *and*-conjunction is overused, but how learners used them in writings and what could explain the overuse phenomenon were left out from their discussions.

Up to this point, a clarification of terminology commonly assigned to *and*-conjunction would be necessary, given *and* appears in research under different names. What I infer from previous studies, Halliday's framework and Celce-Murcia's classification of cohesive conjunctions is the *and*-conjunction can be classified into two types: the *and*-coordinator connects constituents (words, phrases, or clauses) within a sentence, and the *and*-additive conjunction (or *and* as a discourse marker) functions at the supra-sentential level to connect sentences. Some researchers alternatively use "*and*-connector" or "*and*-connective" for "*and*-additive conjunction" even though the words *connectors* and *connectives* have a generic meaning. Thus, I only use the term "*and*-coordinator" to indicate the *and* functioning at the intra-sentential level, and this type of *and* is the focus of my study.

2.2. The acquisition of *and* by ESL learners

What fell short in the common scholarly discussion about *and*-conjunction is picked up by studies about cross-linguistic transfer, which gather evidence on errors or nonnative-like use of L2 to justify or disqualify the hypothesis that L1 influences learners' use of L2. Despite the controversy, transfer theory cannot be refuted because there are demonstrations of how language learners prefer L2 patterns similar to their L1 conventions, how they recourse to L1 at different developmental stages to make a choice in using L2, and how factors other than transfer, such as learners' perception of writing purposes, the target audience, genres, and writing instructions, are found but unable to replace transfer to be the exclusive causes of learners' nonnative-like use of L2.

Among advocates of transfer theory, Radwan (2012), by examining how conjunction *and* in sentence initial position varies among three learner groups—the

intermediate Arabic, the advanced Arabic, and the English native speaker groups—concluded that the higher level of proficiency the Arabic group possesses, the more closely they approach their frequency use of *and* to native speakers'; in other words, the frequency of *and*-initial decreases significantly as the writers' L2 proficiency develops. Although Odlin (1989, p.68) warned against the attribution of transfer to this phenomenon in Arabic writing because of its commonality in other non-native writing, Radwan (2012), as well as Pishghadam and Attaran (2012) and Yorkey (1974), drove home the point: the overuse of *and* by Arabic writers reflects the Arabic writing convention that requires parallel structures and balance of thought; *and* in their writing is used to emphasize and create balance rather than cohesion (Radwan, 2012). Radwan concluded from his study of *and*-initial that as learners evolve into higher stages, they increasingly overcome language transfer from L1, hence the decrease in *and*-initial.

The intertwinement of developmental proficiency and transfer is one of the main interests of the interlanguage hypothesis. Chiu (2004) analyzed the common grammatical error patterns accompanying *and* in writing of three Taiwanese college groups at high-, mid-, and low-proficiency levels; e.g., the omission of *and* or past tense morpheme. Her main study's weakness is the small sample size; therefore, the statistics (percentage) is unable to show how the distribution of errors differs among the groups, but Chiu's summary of error patterns is worth noticing. Even though these errors do not exclusively exist in *and*-structures, they allude to the effect of transfer on learners' decisions, and examples of these decisions relevant to the present study of *and-(*a)symmetry are the omission of *and*-conjunction in complex sentences where *and* should co-occur with the conjunctive adverb *then* (high-level:12.5%, mid: 12.5%, low:11.9%) and the selection of

wrong parts of speech for the conjuncts (high: 9.4%, mid: 8.3%, low: 5.6%). The omission of *and* in the phrase *and then,* according to Chiu, may derive from the interference of the students' first language: Conjunctions in Chinese are not deemed important, and the Chinese phrase equivalent of *then* can be used as either a phrasal or a clausal linking word without *and* before it. The failure to convert the conjuncts (e.g., *curiosity* versus *naughty* in "*Thanks to their curiosity and naughty,..")* into the same part of speech, Chiu noted, is probably because Chinese (Taiwanese) learners are occupied by the meaning of words in the conjuncts and, thus, overlook their part of speech. The numerical findings inquire further investigation because they seemingly suggest that low-level students make fewer mistakes of these types than higher-level groups.

Leung (2005) proposed a similar justification for the difference she found between American and Chinese (Hong Kong) university students' use of *and*-conjunction. The much lower frequency of *and*-coordinators in Chinese writing as compared to American writing may signify the influence of first language transfer: Chinese students carry the way they use conjunctions and connectors in Chinese to the target language (English). It means they use conjunctions or connectors in English at the same rate they do in Chinese. Because Leung considered American students' essays as the reference, she concluded Chinese (Hong Kong) students underuse English conjunctions (which, according to Leung, include coordinators and subordinators) and overuse English connectors (i.e., conjunctive adverbials).

At discourse level, L1 transfer seems to regulate discourse structuring of *and*-constructions. Advanced Norwegian speakers of English, for example, use *and*-coordination in narratives differently from how native speakers of English do in three

main accounts: they employ more S-conjuncts and less embedded complex structures within one conjunct while English speakers more VP conjuncts and more complex units, often involving subordinate clauses integrated in one of the conjuncts, as in "As he probes the sandy earth around him for the source of the water, the earth seems to disintegrate in his grasp and a hole forms where his body occupies and he falls in" (Behrens, 2008, p. 20). In S-conjunction where animate and inanimate subjects are conjoined, they prefer an animate subject for the first conjunct; English native speakers, in contrast, prefer the inanimate. Lastly, the discourse relation between the two conjuncts in Norwegians' writing seems to be interpreted as Background-Action/Reaction (e.g., "***He seems desperate for water*** and he searches in the sand around him for hints of a water source"), but in English native speakers' texts, it has consequential reading or Cause-Consequence and the first S-conjunct is often reduced to syntactically subordinate initial adjunct (e.g., "***Again frantic***, Lofnu put all his strength and effort into enlarging the hole"; or "***He looks around*** and a piece of paper is blown directly in his face") (Behrens, 2008, p. 23).

Explaining the differences, Behrens cited Ramm's study to postulate Norwegian convention of discourse structure transfers to L2 writing. Because Norwegian coordination is less constrained than coordination in English, Norwegian narratives in English involve more Background and Continuation as a discourse relation between the two *and*-conjuncts while the native data indicates a prominent Consequential structure.

From the Systemic Functional Linguistics perspective, *and*-conjunction in L2 developmental writing is analyzed in terms of the parataxis-hypotaxis hypothesis, which posits syntactic moves from "parataxis (coordination) through hypotaxis (subordination)

to ultimate arrival at the extended use of grammatical metaphor (e.g., through nominalization) for achieving the complexity needed to communicate sophisticated notions typical of advanced writers" (Manchón & Norris, 2012, p.235). In other words, proficient learners would use more hypotactic and fewer paratactic structures in their writing, but as a caution, a decrease in hypotaxis could be an indicative of an increase in writing competency.

2.3. Genres' influence on cohesive devices

The relationship between each genre and linguistic properties has resulted in a two-way research course: how genres regulate linguistic traits and how the variation in linguistic features can distinguish genres. The previous is more concerned with the description of genres while the latter raises attention toward genre classification. Most research in the second direction involves a cyclical procedure, in which preliminary description (of topics or other text features) assigns texts to categories, and further investigation to describe these categories may divide texts into sub-genres. Alternatively, if distinctive characteristics of the category are to be found, they can become the yardstick to recognize other similar texts.

Previous research attempting to describe genres has shown that genres can influence the amount of cohesion or the choice of cohesive grammatical devices. Smith and Frawley (1983) found that religious and fiction texts were more conjunctive than science and journalism, as there were much less coordination in the latter genres than in the first. The frequency of cohesive conjunction, however, did not correlate with sentence length in all the three genres except religion, with fiction having the shortest mean sentence length (Smith & Frawley, 1983, p. 6). The authors explained this significant

difference between fiction and science based on two other studies by Gutwinski (1976) and Greenbaum (1969). They found that science is more logic-oriented, hence a greater use of logical sentence adjunction, while fiction was sequential narration by nature. Regarding subordination, science used "*although, since* and *than*" more frequently than fiction, which suggested the comparative nature of cohesive ties. Meanwhile, fiction had the most occurrences of "*before, if,* and *like*," which reflected the analogous and serial characteristics of fictive texts (Smith & Frawley, 1983, p.8)

The description of texts across genres is the foundation of text classification, which can be automated based on knowledge engineering (KE) or a more popular approach called machine learning (ML) (Sebastiani, 2002). ML utilizes an inductive process to recognize (or learn) from a set of preclassified documents the characteristics of interest—characteristics which serve as the crux for an automatic text classifier. ML is more powerful than other content-based techniques, which assort documents in terms of their topics, because it can make use of any feature of texts, such as grammatical structures, to generate a highly accurate result, and thus becoming a popular method for genre classification. Argamon and Dodick (2004) used conjunctions and modals to classify texts into two genres: historical and experimental science writings on paleontology and physical chemistry respectively. The result showed that machine learning techniques trained on data from journals of these two genres achieved over 83% accuracy for the classification tasks. Although Ikonomakis, Kotsiantis, and Tampakas (2005) considered conjunctions stopwords—useless words that would be erased during the process of training classifier, Argamon and Dodick's study shows how the use of

conjunctions and genres are closely related; thus, conjunctions should not be ignored in the machine training process.

In research where texts characterizing and categorizing interlock as recurring steps, much effort is directed toward making use of genre analysis methods to clarify any subtle disparities between texts of the same genre to subdivide them. To that end, researchers combine methods; in examining writings in the BAWE corpus, Nesi and Gardner (2012) drew on Systemic Functional Linguistics (SFL)—known as the Sydney School approach, the SFL appraisal framework proposed by Martin (as cited in Pascual & Unger, 2010), and the multidimensional analysis introduced by Biber (1988). They found that Life Sciences are the most informational; Arts and Humanities involve the greatest amount of narration; Social Sciences are the most elaborated; and Physical Sciences are the most impersonal and persuasive (p.14). Likewise, Gardner (2008) integrated corpus linguistic description of texts, multidimensional analysis of register, ethnographic investigation and SFL to describe genre differences between history and engineering assignments.

The analysis methods in their studies rely mostly on linguistic features, among which is conjunction as an essential descriptor of a genre. For example, to assess the heteroglosssic engagement (or how voices of opposing views communicate), the SFL Appraisal framework—an approach to describe how language is used to adopt stances with respect to the others'—includes concessive and consequential conjuncts as indicators of dialogic contraction (White, 2003; Pascual & Unger, 2010). To measure the degree of impersonality along Dimension *non-impersonal versus abstract and impersonal*, multidimensional analysis takes into account *thus* and *however*—terms that

more frequently appear in written texts than in casual conversations (Nesi & Gardner, 2012, p. 13).

2.4. Genre's influence on coordinating conjunction *and*

Like any other conjunctions, the use of *and*-connective is influenced by genres, but research has yet reached a unanimous conclusion because genres can interact with other features of language production, such as medium (spoken vs. written), to yield different findings. Leech, Rayson, and Wilson (2001), for example, examined four genres in the British National Corpus Sampler: conversational talk, formal speech (task-oriented speech), imaginative writing (e.g., fiction), and informative writing (e.g., news and science). They discovered that, while *and*-coordination is previously considered more common in written language than in spoken language, task-oriented speech has significantly more *and*-coordination than imaginative writing. This finding drew home the point how the results about coordinating conjunctions in speech and writing should be interpreted with discretion.

Within written genres, they observed a higher frequency of *and*-coordination in informative writing than in imaginative writing. This finding seems contradictory to Smith and Frawley's analysis (1983) of 16,000-word samples of four genres (fiction, journalism, religion, and science) in the Brown English Corpus, which demonstrated a lower frequency of *and*-coordination in journalism (300 times) and science (368 times) than in fiction (537 times). The relative frequencies of *and* based on the total coordinating conjunctions in each genre were 77.4% (journalism), 83.8% (science), and 77.6% (fiction), indicating fewer coordinating conjunctions occurring in journalism and science as compared to fiction, but the distribution of *and* in the coordinating-conjunction

category was the highest for science (83.8%). If the within-group distribution of *and* is actually what Leech et al. (2001) referred to when they discussed about *and* occurrence in informative writing (news and science) and imaginative writing (fiction), then a relevant question would be, what can be inferred from these seemingly contradicting numbers to help clarify genre influence on the use of *and*-coordinator.

Greenbaum explained, informative or argumentative texts, such as science, demand a greater use of logical sentence conjunctions than does narrative texts, hence rendering the existence of coordinating words less necessary (as cited in Smith and Frawley, 1983). This hypothesis accounts for *and*-coordination occurring less in science. Science, however, may have lexical content difficult for readers to decipher, so it compensates by seeking cognitively simple structures to facilitate reading processing; thus coordination becomes more prevalent in science texts as compared to other conjunctions, and so does *and* because the statistics show *and* representing the coordination category (Cheong, as cited in Smith & Frawley, 1983). After all, Smith and Frawley did not clarify why *and* is selected over other coordinators; except when mentioning *and* as a cohesive conjunction functioning at inter-sentential level, they attributed rhetorical structures of the text type (genre) to coordination frequency. For instance, science has more additive conjunction *and* than religion, while religion has more contrastive *but* because of the "sequential argument structure in science," and the "assertion [or] contrast structure for religion" (p.353).

With regard to *and*-cohesive conjunction occurring at extra/inter-sentential level and often at sentence-initial position, Bell (2007) compiled a corpus of over one million words from 11 journals representing science, social science, and humanities. He

confirmed the tendency of avoiding sentence-initial *and* (SIA) in writing with the stigma held strongest in scientific writing and far less in humanities journals. In these journals, SIA is the most prevalent additive, and in all journals, SIA occurs three times less than SIB (sentence-initial *but*). For the present research, SIA will not be included due to its distinction in syntactic and semantic behavior as compared to *and*-coordinator.

2.5. Review of methodology in previous research

2.5.1. The qualitative/quantitative dyad and sample size

Substantial research on cohesive devices in writing has long combined both qualitative (close-up investigation of texts) and quantitative (statistical analysis of texts) approaches. For example, to clarify the role of lexical and grammatical cohesion in good and poor compositions, Pritchard (1980) calculated their average use and frequency use in 22 good and 22 poor writings. Finding no distinguish difference between the groups, he took a closer look at texts and realized that unskilled writers often fail to create coherence text with cohesive devices.

Even when the sample size of texts is large, researchers who seek meaningful statistics look into the text for explanation. Hinkel (2001) demonstrates this point with her large sample of 897 academic essays from NS and NNS. The statistics confirmed Arabic speakers employ more coordinating conjunctions than NNs, and Indonesian speakers use less. It is further elaborated in the qualitative discussion, that Arabians' texts maintain rhythmic balance with coordination to add or emphasize writers' point, while Indonesians' texts tend to be repetitive with conjuncts being synonyms or sharing similar ideas. On the other hand, the qualitative section unfolded the syntactic complexity of coordinating structures in NNs' texts, and much short, simpler structures in Indonesians'.

A large sample, however, poses challenges to qualitative description. Although Hinkel (2001) provided a perceptive observation into the text itself, she did not examine other relevant issues, for instance, in what circumstances coordinators are unnecessary or redundant in Arabians' texts, or at what syntactic level Arabic writers most attempt to create balance. Another issue with a large sample is that its benefit to result generalizability could be undermined by many elements: The data may be contrived with linguistically arbitrary restrictions or edited heavily (Barnbrook, Zyngier, & Viana, 2011).

The current research trend obviously favors a combined method for close-up investigation of texts yet meaningful statistics. The question is what sample size this present study should draw on to grant statistical significance testing and qualitative observation.

2.5.2. Bottom-up versus top-down (and/or corpus driven and corpus based)

For long corpus research has employed two main approaches: top-down and bottom-up. According to Haan (2010), the top-down approach is more traditional, starting with a predefined framework or category of the linguistic feature and continuing with an examination of the corpus to segment texts according to the framework. A top-down driven question, for example, could be "to what linguistic category does this feature belong" (p.103). In contrast, the bottom-up approach is exploratory along the way until the researchers successfully deduce characteristics of the linguistic element of interest in the corpus to construct a new theoretical framework.

These two approaches correspond to the difference between *corpus-based* and *corpus-driven*, respectively (Hann, 2010), thus they are often used interchangeably. In

fact, *bottom-up* is more of a general term while *corpus-driven* method is known associated with neo-Firthians who reject the notion of corpus linguistics as a method and insist upon its status as a source of linguistic theories, a new branch of linguistics (McEnery, 2012). They oppose the annotation of corpora because it makes use of extant linguistic features. An exemplar of corpus-driven approach is linear unit grammar (see Sinclair & Mauranen, 2006), which completely abandons the traditional word classes and build a new model based on chunks of text.

The division between c*orpus-based* and *corpus-driven* is, however, what McEnery called "overstated" because to some extent, corpus-driven linguists have to apply their understanding of existing theories to their work (Enery, Xiao, & Tono, p. 10). Thus she dismissed the notion of *corpus-driven* and referred to all corpus research as *corpus-based.* On the other hand, *top-down* and *bottom-up* seems to have a more clear-cut distinction, provided that they are understood as deductive and inductive approaches, respectively. Thus, they can co-exist and interlock. Charles (2007), for instance, used corpus to teach her international graduates, who were working in 27 different research fields at Oxford University, about the rhetorical pattern for defending against counterarguments (as cited in Flowerdew, 2009). The top-down process is manifested in the fact that she explained the rhetorical functions to her students, so they can realize them in texts (e.g., concession). The bottom-up process took place after students identify the rhetorical function as requested. At this stage, her students set out to explore the lexical feature of, let say, concession. They were not told that concessive *while* co-occurs with *acknowledgement* and *see,* but they discovered this lexico-grammartical structure by themselves.

Perhaps the most conspicuous distinction between the two dyads probably lies in the relationship between two terms in one dyad: while *top-down* and *bottom-up* can be combined to analyze corpora, *corpus-based* and *corpus-driven* seems to be either mutual exclusive or very similar to each other.

2.5.3. The approach for this research

Linguistic data from which the sample is drawn may be not normally distributed, and a small sample size (i.e. less than thirty) is heavily dependent on the population distribution, meaning that such parametric tests as ANOVA may be inapplicable because they assume normality of the population. Parametric testing, however, is more powerful than non-parametric. Thus, the present study includes at least thirty texts for each group under research investigation in order for the ANOVA, a type of parametric test, to be performed on the data. This size allows central limit theorem (i.e., assumption for normality regardless of the population distribution) to take effect, so that the data can be analyzed by statistical significance testing. At the same time, a non-parametric test equivalent to ANOVA was utilized to cover the possibility of a non-normal population.

Gries criticized far-reaching claims in terms of frequent phenomenon based on a few examples (Barnbrook et al., 2011). In his opinion, even 200 examples are not enough. The present research assumes that a sample of 30 texts, each of which has more than 500 words, is enough to make claims because "the size of corpus needed to be representative is closely linked to the frequency of the items(s) under examination" (Leech et al., 2001). Given the considerably high frequency rate of *and* in texts, instances of *and* in total would be large.

The research is corpus-based, top-down oriented since the (a)symmetry feature of *and*-coordination is classified based on the existing grammatical category or at least derived from it, and the corpora (i.e., MICUSP and NUCLE) from which the samples are selected are annotated corpora.

This study also takes into account the influence of genres, and, therefore, limits the genre (text type) to argumentative essays. The study, however, has a main weakness due to not controlling the disciplines and topics of the essays. A detail discussion of the essays' disciplines and topics will be presented in chapter 4, "Methodology."

Chapter 3: Literature review of *and*-coordination, the research question, and the hypotheses

3.1. Definition and notable features of the coordinating conjunction 'AND'

According to table 1, simple coordinating conjunctions include *and, or, nor, so, but, yet,* and *for.* They conjoin two constituents with the same grammatical function or status (Celce-Murcia & Larsen-Freeman, 1999). This study concentrates only on the coordinator *and,* the most common and probably the most overlooked conjunction in EFL research. The definition of *and* as joining *the same or equal* elements appears to be simplistic and hence require a encompassing review to account for many linguistic instances of *and* that seems contradictory to the definition. The trajectory to which my research adheres stems from the theme of syntactic and semantic symmetry and asymmetry accompanying *and.*

One comprehensive summary of the characteristics of *and* is written by Celce-Murcia and Larsen-Freeman (1999). *And,* at its basic function, overtly signals the *plus* cohesive relationship between even-level items, such as those of the same part of speech (N plus N, and V plus V), the same phrase type (PrepP plus PrepP), and the same clausal or sentential structure, to create a larger constituent of the same type. The conjunction can connect more than two items, and when it is the case, the *and* between the last two items is retained while all of its other instances are deleted. The author noticed, the conjuncts—the linked constituents—have distinctive, yet analogical meanings; for example, *neatly* and *effectively* refer to different manners of an action, but they share the positive connotation, as in "The problem was solved neatly and effectively" (p.462). The

similarity in syntactic structures of the conjuncts creates what Blakemore (2005) called *syntactic symmetry*.

Yet the authors' analysis of *and* expands to three syntactic complications, which happen when conjuncts are clauses: ellipsis, pro-forms substitution, and gapping. Ellipsis rules require the omission and replacement of the verb phrase in the second conjunct with a compatible auxiliary, and the simultaneous addition of the adverbs *too, so, either,* or *neither.* In some cases, ellipsis co-occurs with pro-form substitution whereby pronouns and substitute words such as *one, then,* and *there*, are used in place of repeated words. Gapping is a special case of ellipsis: the elided part of the sentence lies in the middle rather than at the end of the second clause, provided that conjoined sentences have "non-identical subject" and "at least one non-identical predicate constituent apart from the verb" (Celce-Murcia & Larsen-Freeman, 1999, p.469). An example of gapping is "John trimmed the tree, and Marry the head." Whether the structures of these three complications are symmetric or not will be discussed later in this section.

In terms of meaning, Celce-Murcia and Larsen-Freeman acknowledged that *and* conveys more than the semantic symmetry *plus.* In the example "Fred fell down, and he hurt his foot badly," the word *and* implies a cause-result connection; therefore readers would understand the sentence differently if the two clauses interchange. The authors cited a list of possible meanings of *and* outlined by Posner (1980), including sequential ("Peter married Annie, and she had a baby") and hypothetical ("Give me your picture, and I'll give you mine") cohesion (p.474). This complexity queried the exact meaning of *and*. But Blakemore's explanation seems to be the resolve Celce-Murcia and Larsen-Freeman was seeking, because Blakemore separated the actual meaning and the aspect of

use: "…these so-called 'meaning' of *and* are actually not meaning at all but aspects of use—where 'use' means that the context of utterance determines exactly how the word will be interpreted" (as cited in Celce-Murcia & Larsen-Freeman, p.474).

A key question of my study is whether relying on Celce-Murcia and Larsen-Freeman's summary alone suffices to accommodate how many categories into which the instances of *and* can be classified. Recent research clustering around the topic of (a)symmetry of *and* has been few. A 2005 *Lingua* issue, edited by Blakemore, was devoted to a highly theory-driven, on-going discussion about the topic of coordination's (a)symmetry, shedding light on my issue of interest from the Minimalist perspective, and together with Butler's work (2003), provided a multidimensional view on *and*-coordination.

3.2. *And*-coordinaton from the Minimalist Program and SFL perspectives

3.2.1. Minimalist Program

Artstein (2005) explored the semantic ambiguity of the conjuncts caused by elision in the structure of the conjuncts. The meaning of *and* is equivalent to *plus*. The type of ambiguity Artstein mentioned comes from plural morphology that refers to a cumulative plural relation rather than an individual relation, as in "Bill and Martha are ortho and periodontists." The question that the study resolved is which word parts that the conjunction *and* operates on to have this result.

Utilizing the semantics of *respectively*, however, could dissolve the ambiguity and make the sentence grammatically correct: "Bill and Martha are an ortho and periodontists, respectively."

His study parallels Celce-Murcia and Larsen-Freeman's conclusion on the case of ambiguity at the level of sentence and clause (Celce-Murcia and Larsen-Freeman, 1999, p. 470). As inferring from Artstein's discussion, an *and*-coordinator operates on one-level below the level of a full structure. For example, in the equivocal sentence "Mike and Jim are husbands of Mary and Jill," *and* connects two elements below the predicate level (Artstein, p. 360). Its full (but not equivalent) form, "Mike and Jim are husbands of Mary and husbands of Jill," has *and*-coordination between two predicates. Similarly, the sentence "John and Mary got married," which cannot be derived from a full, sentence-coordinated form "John got married and Mary got married," is an example of phrasal conjunction. This information is useful for classifying other similar ambiguous structures in my study.

To sum up, Artstein's study drew attention to four cases: Subject = plural NPs with *and*-coordination (Bill and Martha); S = singular NP (Bill); Predicate = plural NPs (ortho and periodontists); and Predicate = singular NPs (an ortho and periodontist). The possible combinations and the results are as follows:

Subject (singular NP) + Predicate (singular NPs) → no ambiguity

Subject (plural NPs) + Predicate (plural NPs) → ambiguity

Other notable findings beyond my research scope are the phonological constraints on coordinate structures, the ungrammatical separation of compositional morphological words (e.g., we need new black and floorboards), and the inquiry on part of words which functions as independent elements in coordinate structures but not in other grammatical constructions.

In another study, Velve (2005) dealt exclusively with the right and left-edge coordinate ellipsis in English with a reference to other languages as well; other types of ellipsis, such as VP ellipsis and gapping, are not considered. Right-edge ellipsis (Right Node Raising) occurs in the initial conjunct(s) before the conjunction, e.g., Pierre bought and Paul read the book (p. 485). Left-edge ellipsis refers to the absence of elements on the left edge of post-initial conjunct(s); such element can be the subject (S), as in "During the lecture something strange came over him and threw him off balance" (p. 486). The author's major attempt is to explain why coordinate ellipsis can occur at certain syntactic positions, i.e., what factors constrain and license the deletion.

Te Velve's proposal expanded on Celce-Murcia and Larsen-Freeman's coordination ellipsis. The main difference is, he adopted the Minimalist framework within which the X-bar theory of Generative Grammar is abandoned, to assume the syntactic asymmetry for coordinate structures, illustrated by the conjunction-as-a-head phrase schema (Conjunction P). Such syntactic asymmetry is magnified in ellipsis phenomena, but at the same time, te Velve conceded that "symmetry is essential to ellipsis and plays an important role in [&] licensing a gap" (p.499). By proposing that, for the coordinate ellipsis to occur, the syntactically symmetry should be set up, he allowed his theory to intersect the Generative Grammar approach: the conjuncts are syntactically symmetrical for they belong to the same syntactic category.

Regarding *and*-coordination in spoken discourse, Blakemore and Carston (2005) paid attention to the semantic (a)symmetry derived from the maxim-driven pragmatic inference—inference of what the speaker means in a specific communicative situation based on general norms called pragmatic maxims (or in other words, how people

normally respond in that situation). Unlike their previous study, discussed by Celce-Murcia and Larsen-Freeman, about *and*-conjunction and its signaling of such logical relations as temporal and causal beyond the simple *plus* meaning, Blakemore and Carston explored another set of *and*-conjunction utterances independent of narrative function, by making use of the relevance-theoretic framework to interpret *and*-utterances that do not convey temporal and causal relations. Their relevance-theoretic framework for *and*-utterances states that listeners infer the meaning behind the utterances by recognizing the relevance of each separate *and*-conjunct to the communicative situation and at the same time processing the total effect of the conjoined conjuncts. The authors' long-term goal is to establish a consistent pragmatic account of all the *and*-utterances in speech.

This special set of *and*-utterances was explored in argumentation cases when the speaker intends to argue against the other's assumption, hence the irrelevance of the possible sequential or causal relations drawing from the positions of the conjuncts. In fact, their study's examples, such as "John broke his leg and he tripped on a Persian rug," defy logical deduction, but in a conversation where the context and intonation clarify the speaker's intention, this seemingly-illogical sequence still maintain coherence. Similarly, in attitudinal cases where speakers intend to show their surprise to the state of affairs, the reversal of the conjuncts makes no difference to the meaning of the whole utterance. But again, according to the relevance-theoretical framework, when listeners take into account the meaning of both conjuncts, they realize these *and*-utterances constitute some sort of contrast or conflict between the states of affairs depicted in the conjuncts, as in the example, "Her husband is in the hospital and she is seeing other men."

3.2.2. SFL perspectives and Functional Grammar

Butler (2003) drew on the Functional Grammar (FG) definition of coordination—a definition postulated by Dik (1997): "A coordination is a construction consisting of two or more members which are functionally equivalent, bound together at the same level of structure by means of a linking device" (p. 184). In FG, the conjuncts are recognized locally, "at a number of places and levels of underlying structure," in order for FG linguists to circumvent the ellipsis rules projected by transformational grammar (TG), to account for both simple coordination (single constituents) and the simultaneous coordination (pairs and larger groupings). They consider cases, such as "Carolyn cleaned and Martin tidied the cutlery board," the coordination of [subject + predicate], instead of [predicate + object], so the deletion rules do not apply.

The simple coordination FG delineates with a focus on clause can be between independent clauses ("I'm a divorced forty year old woman with two grown up children *and* I am in a terrible situation"), between arguments under a verbal predicate ("But a minister said that politics did not work that way, *and* that the influence of the East Anglian connection was greatly exaggerated"), and between clauses that have the same, or sometimes, different illocutions, as in the following example in which a declarative is coordinated with an interrogative, "…modern apprenticeships are going to help that process even further *and* why aren't modern apprenticeships being welcomed?"

In tackling "Verb Phrase" coordination, FG dismisses the category VP; thus, its alternative explanation for such cases as "Blanche waited and watched him intently…" is the existence of a zero anaphora in the first argument position of the second clause. In analyzing simultaneous coordination (or what TG refers to as *gapping),* FG calls for the use of coordinated n-tuples of slots, each of which comprises a human Agent and a Goal,

with the contingency that each of the Agents, as well as each of the Goals, is different. For example, the gapped sentence "…she played the schoolgirl and he a teacher" can be analyzed as follows: *play[V]* $\{(x_1{:}\langle hum\rangle)_{Ag}\ (x_2)_{Go}\}$ & $\{(x_1^{*}{:}\langle hum\rangle)_{Ag}\ (x_2^{*})_{Go}\}$ (The slot is $\{(x_1{:}\langle hum\rangle)_{Ag}\ (x_2)_{Go}\}$).

So far I have discussed the definition of *and*-conjunction in terms of its syntactic, semantic, and pragmatic functions, approached by Transformational Grammar (TG) and Functional Grammar (FG). Within TG, General Grammar linguists (Celce-Murcia and Larsen-Freeman) rely on the X-bar theory in which *and*-coordination is an exception, to anatomize *and*-coordinating structures, while Minimalists (Arstein, Blakemore, Carston, Cormack, and Velve) advocate an asymmetric tree structure with conjunction as the head (Conjunction Phrase). Both schools recognize the ellipsis phenomenon; in contrast, FG does not, and analyzes linguistic expressions in terms of an abstract underlying predication. Despite the discrepancy, both TG and FG assume symmetry at least in the sense of equal syntactic or structural status of *and*-conjuncts.

The underlying *and*-symmetry to both syntax and meaning discussed by TG and FG linguists, from my vantage point, should not be the reason to ignore any possible syntactic asymmetry caused by structural complications or differences of the conjuncts (such as *Rise Node Raising*), and semantic asymmetry by truth-conditionally in-equivalence (i.e. X&Y # Y&X), because they may influence the way learners, especially English language learners, use *and*-coordinate structures. On the one hand, learners may make mistakes when it comes to complicated structures. On the other hand, they may avoid those structures as much as possible.

For this study, a combination of TG and FG is necessary to predict and categorize as many *and*-instances that may arise during the text analysis procedure. I suggest with recourse to both TG and FG **the working definition** of *and*-coordination is, *and*-coordination is the process of combining two or more members of the same level of syntactic status within a sentence, to produce a larger structure of the same level of syntactic status, by means of the linking device – functional word *and.* This definition improves on Celce-Murcia and Larsen-Freeman's account of coordination, "…coordination is the process of combining two constituents of the same type to produce another larger constituent of the same type," to accommodate the specific case of *and*, and to clarify what it means by *of the same type* in their definition. In addition, it refers to the surface syntactic symmetry as *the same level of syntactic status*, leaving room for the possibility of syntactic asymmetry: Two members having the same level of syntactic status can be not the same type. For example, in *We love red beans and rice*, the word *and* coordinate two noun phrases, *red beans* and *rice*, within the verb phrase *love red beans and rice* of the predicate, but the noun phrases are not quite the same type: *red beans* consists of the adjective phrase (AP) *red* modifying the noun phrase *beans* while *rice* contains no AP.

Based on this definition, I categorize *and*-coordination into two groups of syntax: symmetric and asymmetric. Syntactic symmetry refers to coordinated conjuncts that are very similar in syntactic structure, such as those in category one in table 3. Syntactic asymmetry refers to coordinated conjuncts that somewhat differ from each other in term of syntax, e.g., affirmative versus negative, imperative versus affirmative, copular verb versus auxiliary verb, inverted versus uninverted, elided form (with gap in the middle or

at the end) versus full form, and transitive verb versus intransitive verb. The list may be not exhaustive, but exemplify the underlying reason for the categorizing decision. The examples in table 3 are either collected from Butler (2003), Celce-Murcia and Larsen-Freeman's summary (1999), and Velve (2005), or self-created.

Table 3

Categorization of and-conjunction with regard to syntax

		Symmetric	Asymmetric
Category 1: Similar conjuncts	N+N	Bread and butter	Defense and improving weaponry
	NP+NP	The bread and the butter	
	Adj+Adj	Big and strong	
	AP+AP	Very big and extremely strong	
	Verb+verb	Quickly run and hide	
	VP+VP	Run fast and hide quickly	You can eat those herbs and not get cancer.
			This method is time consuming and does not make economical sense to businesses.
	PrepP+PrepP	Over the field and into the trees	
	Adverb+adverb	Neatly and effectively	
	AdvP+AdvP	Very neatly and rather effectively	
	S+S	Juanita is brilliant, and Shalimar is personable.	She got into the pool, and she began to swim. Do that and you're guaranteed 20-plus goals. [M]odern apprenticeships are going to help that process even further *and* why aren't modern apprenticeships being welcomed?
	CP+CP (complimentizer phrase)	[A] minister said that politics did not work that way, and that the	[A] minister said that politics did not work that way, and that the

		influence of the East Anglian connection did not improve the situation.	influence of the East Anglian connection was greatly exaggerated
Category 2: Ellipsis with *too, so, either,* and *neither*	Uninverted affirmative form	Birds can fly, and I can too.	
	Inverted affirmative form		Birds can fly, and so can I.
	Uninverted negative form	She hasn't left, and I haven't either.	
	Inverted negative form		She hasn't left, and neither have I.
Category 3: Ellipsis with gap	Gapping		John trimmed the tree, and Marry the head.
	Right Node Raising		Pierre bought and Paul read the book.
	Left-edge[a]	Something strange came over him and threw him off balance[b].	Blanche waited and watched him intently.
Category 4: Pro-forms		She has left the country, and I have left it too.	
		He graduated from Tulane, and she graduated from there too.	

Note. [a]Regarding VP+VP category 1, if the head verb in VP is also the main verb in the predicate, the VP+VP structure will be considered as left-edge ellipsis. [b]I decided the example should be symmetry because it seems very much similar to VP+VP in category 1.

Regarding semantics, I categorize *and*-coordination into two groups (semantic symmetry and semantic asymmetry) based on Blakemore's theory (Table 4). *And*-semantic (a)symmetry is, however, outside the scope of this research.

Table 4

Categorization of and-conjunction with regard to semantics

Categories and Explanation		**Symmetric**	**Asymmetric**
Category 1: Truth-conditionally	(X& Y) has the same meaning as	Tokyo is the capital of Japan, and Paris	

equivalent	(Y&X)	is the capital of France (Celce Murcia & Larsen-Freeman, 1999).	
Category 2: Truth-conditionally inequivalent	(X& Y) does not have the same meaning as (Y&X)		Fred fell down, and he hurt is foot badly (Celce Murcia & Larsen-Freeman, 1999).
Category 3: Ambiguity	Includes cases similar to Blakemore's *and-utterances* with argumentative and attitudinal intent. In written texts, non-verbal cues are not available, and if there is no other cues within the text to signal argumentative and attitudinal intents, swapping conjuncts would not create a conspicuous difference.	Paul is a linguist and he can't spell (Blakemore, 2005)	
	Another ambiguity shows in an example in Cormack's study for simple conjunction, but the example can be interpreted as a temporal ordering as well. If the context is insufficient in clarifying the ambiguity, I suggest to classify such ambiguous instance as semantic symmetry, ignoring the meaning of temporal ordering.	He did some weeding and wrote a few pages of the paper (Cormack, 2005).	

3.3. A preliminary analysis of errors with *and*-coordination in the NUCLE corpus

The purpose of this error analysis is to identify evidence that might help formulate the direction (expectation) in the hypotheses. In the NUCLE corpus, transition errors in which learners incorrectly omit *and* or incorrectly use other substitutes for *and* are 296 instances, which make up 20.83% of total 1421 transition errors (the use of wrong linking words or phrases). This preliminary error analysis does not count errors, if exist, in cases where *and* should not have been used. The total number of *and*-coordinators learners use excluding those in the error-tags are 25,835.

A quick examination on the first one hundred instances of errors with *and*-coordination among a total of 296 in the NUCLE corpus showed the following noticeable findings:

a. 74% of the errors are related to ***semantically*** *s*ymmetric coordination. Among those 74 instances, 55 are the omission of *and* in item listing and clausal connection when *and* is necessary (74.3%). Other errors include confusing *and*-coordinator with relative pronouns *which* and *that* (8.1%), and with other coordinators such as *but*, *or*, and *for* (10.8%). It is not clear why students have more errors related to semantically symmetric coordination than semantically asymmetric coordination. It might be simply because they use more semantically symmetric than asymmetric coordination. In cases where they confuse *and* with other coordinators, they seem to have difficulty with determining the semantic relationship between conjuncts.

b. Even though confusion with relative pronouns only accounts for 8.1% among 74 errors, 14.9% actually appears in structures containing relative clauses or reduced forms. Learners seem to have problem with deciding which elements the relative pronouns

modify, and they seem to assign dependency of the state of affairs in one conjunct to that of the other conjunct instead of a parallel relationship.

c. There are slightly more errors (i.e. omission of *and* or incorrect substitution of *and* with other words) present in ***syntactically*** asymmetric structures (55% of the total errors with *and*) than those in syntactically symmetric structures. It might be that the perception of syntactic asymmetry influences the use of *and*-coordinator. Based on the statistics in b) and c), I suspect syntactically asymmetric *and* might be a problematic area for ESL learners in the NUCLE corpus. The avoidance phenomenon in which learners avoid using certain target linguistic elements because of their difficulty (Gluth, 2003) is likely to occur in this area. The speculation helps formulate the hypotheses in section 3.5.

3.4. Justification of the research

This research examines *and*-coordinator because *and* is among the top frequently used conjunctions, but has long been overlooked because of its seemingly simple structures. A limited number of studies have touched on issues related to *and* as an additive conjunction, such as *and*-initials used by Arabic learners, *and*-initials in academic writings, the overuse of *and*-connectives in learners' writing, and the frequency use of *and*-connectives across domains. *And* as a coordinator receives more attention in the studies of Arabic learners' writing, but in general, it seems to be assumed not a grammatical area that learners may have difficulties with and that creates much difference in writing among learners. However, my preliminary review of errors with *and-coordination* in the NUCLE corpus suggests otherwise, that learners do make mistakes in using *and*, and the proportion of those errors to the total transition errors are

not small. Given previous research findings and the above error analysis, the question is does language proficiency influence the use of *and*-coordinator and in what way?

There is not an easy answer to the question. Research has not offered a well-rounded picture, so it is essential to contribute a piece to this jigsaw puzzle. The present study speculates that a potential answer might reside in the structure of *and* itself. Learners have to make sure they select similar elements for the conjuncts when they use *and*-coordination. But parallel conjuncts do not always constitute a perfect symmetric nature. Their asymmetry in terms of syntax permits coordination of elements of the same syntactic level but belonging to different grammatical categories (i.e., auxiliary verb versus ordinary verb), which evokes more thought from learners to transform deep structures into grammatically correct surface structures. It elicits more effort to determine which elements are parallel and should be conjoined by *and.*

The thesis has a scientific and a practical relevance. The scientific relevance is the contribution to clarify learners' second language acquisition of *and*-coordinator, namely, whether and in what way language proficiency affects the use of *and*-(a)symmetric structures, especially in academic writing. It expands the seemingly unanimous conclusion that *and*-connective is overused by learners by analyzing the overlooked type of *and*—the *and*-coordinator—not only in its superficial symmetry but also in its deeper asymmetric structures. The practical relevance is the pedagogical implication. If the hypotheses are supported, English teachers may find it necessary to draw learners' attention to the syntactic (a)symmetry of *and*-coordination, and what alternative choices they can have to replace *and*-coordination, including using other cohesive conjunctions and grammatical forms.

3.5. Research questions and hypotheses

My research questions concern the differences in frequency of *and-coordination* in argumentative academic writings among the three learner groups: English native speakers at the advanced level of academic knowledge, non-native speakers of English at the advanced level of academic knowledge, and non-native speakers of English at the upper-intermediate level of English proficiency.

The dependent variable is the frequency of *and-coordination* in argumentative academic writings. The independent variable is learners' English proficiency (NSs at the advanced level, NNSs at the advanced level, and NNSs at the upper-intermediate level). What constitutes English proficiency will be discussed in section 4.2.3, chapter 4. The stratification of the independent variable is based on the assumption that English native speakers at high levels of academic knowledge would exude the highest English proficiency, and non-native speakers of English at lower levels of academic knowledge and proficiency would demonstrate the lowest English proficiency among the three groups.

The expectations in the hypotheses are influenced by findings in previous research that *and* is often overused by English learners (only Leung's study found the *and*-coordination is underused); and by the preliminary error analysis of the NUCLE corpus, which points to the possibility that NUCLE writers have issues with syntactically asymmetric *and*-coordination, hence a possible avoidance of this structure.

The hypotheses are as follows:

1. There is a significant difference in means frequency of *and-coordination* in argumentative academic writings among the three learner groups, and the NUCLE group uses *and* the most frequently.
2. There is a significant difference in means frequency of *and-coordination* in terms of its syntactic symmetry in argumentative academic writings among the three learner groups, and the NUCLE-NNS group uses *and* the most frequently.
3. There is a significant difference in means frequency of *and-coordination* in terms of its syntactic asymmetry in argumentative academic writings among the three learner groups, and the MICUSP-NS group uses *and* the most frequently.

The null hypothesis for each hypothesis above is that at least one of the population means is different from the others.

Chapter 4: Methodology

4.1. Research design

The study follows the experimental design in which subjects (essays) were randomly assigned into three groups or three learner corpora. A random selection of 31 texts from each group ensures the representativeness of these samples. A discussion on the corpora as well as their compatibility will be detailed in section 4.2, "Population and Sample." In addition, the writings are independent of each other since they are about different academic topics, and each participant belongs to only one group.

4.2. Population and sample

4.2.1. Michigan Corpus of Upper-Level Student Papers (MICUSP)

As mentioned in chapter 1, MICUSP is a developmental language corpus, having 2.6. million words from 829 A grade papers across sixteen disciplines. The project submission process started in late 2004 throughout 2006, and the data collected was mostly from online submission (by emails and website). 82% of student writers are native speakers of English, leaving the remaining 18% to be non-native writers yet accounting for 148 papers. The L1 of non-native writers for argumentative essays (31 in total) are mostly Chinese (12 students), Korean (4 students), Japanese (3 students); the rest are Tamil (2 students), Thai, Hebrew, Urdu, Bosnian, Bengali, Greek, German, Spanish, and Swedish.

The representativeness of the corpus seems to be fair, with students' submission reflecting the composition of enrollment, gender (62.1% were written by women and 37.9% by men), and the number of writing assignments across fields of study (Römer,

2011). The distribution of papers in the corpus shows the highest submission rate is for the Social Sciences and the Humanities and Arts divisions, and the lowest for "hard" sciences such as the Physical Sciences. Römer suspected an explanation for this imbalance being fewer students enrolled in the Physical Sciences, or fewer written assignments given to the students in the sciences. The length (word count) of the paper is also the highest in the first two discipline groups, and their overall word counts are as high as 978,254 and 734,437, respectively, followed by the Biological and Health Sciences (511,550), and the lowest rate is for the Physical Sciences (329,288). Within the Social Sciences and the Humanities and Arts divisions, there were more female students submitting their papers than men, alluding to the fact that more women enrolled in these fields. The rate of submission decreases across four levels (final undergraduate and the first three years as graduate), which again probably mirrors the enrollment rate.

Texts in the corpus were categorized by their authors and the research team (Römer, 2011). The author of a paper classified his or her paper at the time of submission, but the researchers were those who made the final decision. Argumentative essays, the focus of the present study, are those qualifying the team's consensus about rhetorical purposes and features (table 5) —reflecting Pattridge's definition of genre discussed in chapter 1.

Table 5

Criteria for argumentative essays (Römer, p.170, 2011)

Rhetorical purposes	Features	Examples
Demonstrates ability to construct a coherent argument and support it with evidence /examples	- paper is thesis driven - author's thesis is supported by pieces of evidence from an outside source - may generate a new idea/argument in the field	Argumentative essay, persuasive essay, literary analysis essay

4.2.2. NUS Corpus of Learner English (NUCLE)

The NUCLE corpus is an annotated, error-tagged learner corpus. According to Dahlmeier (2011), the NUCLE corpus consists of over one million words, or about 1,400 essays written by ESL undergraduates in academic writing courses. The essays cover a wide range of topics, including environmental pollution, healthcare, and engineering. In a private correspondence, the author believed the genre to be argumentative essay writing, and some essay topics can be tailored for a specific disciplinary course (e.g., engineering) (email, January 1, 2013). They were take-home assignments with specific question prompts given by the teachers (email, January 31, 2013).

The students were non-native speakers whose L1 were predominantly Chinese and a few were South-East Asian languages such as Malay, Tamil, and Vietnamese. They had relatively high proficiency in English, evidenced by a considerably lower rate of making mistakes on the articles (only 1.8%) and the prepositions (1.3%) (Leacock et. al., 2010). Further detail about the corpus is being documented by the author and, thus, has not been released.

4.2.3. Definition of learners of high English proficiency in academic writing

The concept of learners of high English competence, or *advanced learners*, varies from one study to another. In his dissertation, Guo (2006) expresses concern about a lack of linguistic standards to scale the level of learner English. Apparently researchers set their own criteria for assessment of language production. White (2002), for example, determined advanced Turkish learners of English to be those who achieve above 93 percent on a university placement test. Granger (1998) referred to third- and final-year non-native students as advanced learners (cited in Guo, 2006). Guo introduces Lorenz's

proposal (1999) by which *advanced learners* are those "who meet advanced foreign language requirement, i.e. learners who are generally expected to have mastered the basic rules and regularities of the language they are learning" (p.10). He concludes

> No matter whether theoretically or empirically, there is a need to establish a relative norm so that when someone mentions "advanced learners", it will be explicitly understood as the same (or approximately the same) thing with the same parameters in the measurement of the learners' English. (p.44)

His criticism addresses the common toleration of the inconsistency in evaluating language competency in the field of second language research. Without an available norm, I estimated the English competence of learners in MICUSP and NUCLE via their essays' quality and level of education. The native speakers' writing in MICUSP, I assumed, manifest the highest level of English proficiency because, in addition to their *nativeness* in using English, the writers also achieved grade A for their writing at a high level of their tertiary education (final year undergraduate and post graduate). The writing by non-native speakers in MICUSP can be considered as approaching the MICUSP native speakers' proficiency. The English proficiency of both groups fall on the high end of the scale.

The writing in NUCLE can be considered beyond the intermediate level but below the advanced level. Students who attend academic writing courses (e.g., NUCLE learners), according to the National University of Singapore website, must previously sit for the Qualifying English Test (QET)—a placement test upon admission to the undergraduate program—and achieve a passing score that allows them to skip basic, general English courses (band 1 status). Otherwise, they have to pass those courses

before attending academic writing classes (band 2 status). Only students who have proof of proficiency equivalent to a Toefl-iBT score of 115 or higher can be exempted from the QET, and those who take the test and are assigned band 3 status can skip both general and academic English courses. The description of band 1 status seems to reflect learners' abilities at the intermediate level. The NUCLE students are at a higher level than the intermediate level because they held band 2 status, and this evaluation is also similar to Dahlmeier's opinion. From my observation, the essays in NUCLE demonstrate an ability (although limited) to organize and connect several paragraphs, develop a thesis, incorporate evidence, and integrate in-text citations and references. These qualities are absent in the descriptors of band 1 status. The NUCLE essays, however, are below the proficiency level of the MICUSP non-native writers' essays because the NUCLE writers were assigned to academic classes for one main reason: QET results suggested the writers have difficulty with academic writing, and they needed the course to learn about and improve academic writing skills.

The lack of a norm for assessing learners' level notwithstanding, a reference to available descriptors for advanced learners' writing to supplement my own evaluation would yield more solid judgment. Thus, I consulted the *ACTFL Proficiency Guidelines 2012* developed by American Council on The Teaching of Foreign Languages (ACTFL), to determine the English competency level for MICUSP's and NUCLE's learners. I particularly selected the Guidelines because of their unbiased description and global application. According to their general preface

> The Guidelines are not based on any particular theory, pedagogical method, or educational curriculum. They neither describe how an individual learns a

> language nor prescribe how an individual should learn a language, and they should not be used for such purposes. They are an instrument for the evaluation of functional language ability. (p.3)

ACTFL emphasizes "the Guidelines are intended to be used for global assessment in academic and workplace settings" (p.3). Given MICUSP contains essays for an American university and NUCLE for a Singaporean one, it is reasonable to adopt ACTFL's guidelines.

The Guidelines stratify proficiency on a continuum of five major levels: Distinguished, Superior, Advanced, Intermediate, and Novice, with the latter three further divided into High, Mid, and Low sublevels. Learners at the Intermediate level of writing are characterized by the ability to write for practical needs, including letters and requests. Their writing is filled with simple facts and ideas, loosely connected sentences, and the present tense. Although they can communicate topics of their interest, the vocabulary and grammatical structures they use are basic, and they may have difficulty making themselves understood.

The ACTFL specifies writers at the Advanced Low sublevel are able to meet basic work and academic writing needs. They can summarize familiar topics, compose essays of several paragraphs in length, and incorporate limited types of cohesive devices. Their writing, however, displays redundancy, unskilled repetition, influence of oral discourse structures and writing style of their first language, and attempt at using (but a lack of control of) grammatical structures and vocabulary related to very high advanced levels. Advanced-Mid writers demonstrate more control of the target language. They utilize various cohesive devices to cohere ideas and paragraphs. They also develop ideas

in more detail and clarity. Despite their progress, they have difficulty communicating complex issues or performing tasks typical of Distinguished and Superior levels. The Advanced-High level is associated with the capability to correspond in appropriate conventions (e.g., formal versus informal), handle general and special topics with great precision, and apply various grammatical structures and vocabulary. At times, the writing at this level may show influence of first languages (e.g., organization) and errors, but in general, it can be understood with ease by native speakers.

A reference to the ACTFL's Guidelines as an intermediary for comparing the two corpora would address Guo's criticism about the lack of a common proficiency scale. Because the compiler of the NUCLE corpus has not conducted any formal proficiency assessment of the essays, the ACTFL becomes useful in that it would provide a relative perception about the proficiency level of the NUCLE and MICUSP writing. The NUCLE corpus demonstrates descriptions of the Advanced-Low to Advanced-Mid levels in the ACTFL's Guidelines. The non-native section in the MICUSP corpus seems to fit descriptions of the high end of the Advanced-High to Distinguished levels. I will, henceforth, refer to writing in NUCLE as being at the *advance-low/advanced-mid* or *upper-intermediate* level and to writing in MICUSP (native and non-native) as being at the *advanced* level. It should be noticed that the general level of English proficiency is commonly decided based upon a combination of four skills, namely, reading, listening, writing, and speaking, with linguistic competency (grammar, vocabulary, and pronunciation) included in the description of each skill. Thus, strictly speaking, the level of English proficiency under the discussion is restricted to writing.

Because the essays in my current study are academic-oriented, I do not exclude native students from the definition of *advanced* learners, despite of the more common definition that learners in corpora are meant to be foreign language learners. As briefly mentioned in section 1.4.3—"Developmental Language Corpora"—native students are still in the progress of developing their language in their field of study. Considering them as *learners* implies a comparison between learners' writing, no matter whether those *learners* are native or non-native speakers, and experts' writing even though experts' examples were not included in my research. My viewpoint echoes Gabrielatos and McEnery's distinguishing of two dimensions for comparison: nativeness and expertise (2005).

When *nativeness* is the main criteria of distinction, the language produced by native speakers becomes the norm to which non-native speakers are compared, and, therefore, the comparison would involve two groups, one being non-native learners and the other being native speakers, regardless of whether they are native students or native experts (Gabrielatos & McEnery, 2005). The language for comparison in this case should be the general target language (i.e., general English). When expertise is considered, the language generated by experts in the field of study functions as the norm for division: one group would be learners regardless they are native or non-native students, and the other group would be experts. The language for comparison should be the target language associated with specific disciplines (i.e., academic English). Even though native students may need to learn about academic English, they obviously have advantages over non-native speakers because academic English is not so different from general English. Thus,

native students' essays in MICUSP in this thesis are considered having the highest proficiency level among the three corpora.

Römer (2005) concludes from her studies about nativeness and expertise that in advanced-level academic writing, expertise is more likely a decisive factor for distinction between essays than nativeness. Römer quotes Swales on this issue, "the difficulties typically experienced by NNS academics in writing English are (certain mechanics such as article usage aside) *au fond* pretty similar to those typically experienced by native speakers" (p. 99). If Römer's observation is applicable in most cases, then my study of *and*-(a)symmetric structures might reveal no significant difference between the two subgroups (native versus non-native) in the MICUSP corpus. I will return to this issue in chapter 6, "Discussion of Results, Contributions, and Limitations." Due to my choice of mixed level subgroups sampled from the NUCLE and MICUSP corpora of *academic writing*, expertise and nativeness—indicators of English proficiency (academic and general English)—would influence the results.

An expansion of my discussion above is experts' academic writing should be the norm, not native students' in the MICUSP corpus. With the incorporation of native students into the *learner* category, my study conforms to Bolton et al.'s advocate of experts' writing (i.e., writing found in peer-reviewed journals) as the target model for ELL learners (2002). The concepts "overuse" and "underuse" by non-native speakers, as a result, are not applicable to the present study of *and*-coordination when taking into consideration that three college groups under the investigation are all *learner* groups.

4.2.4. Samples

93 written argumentative texts were collected from the MICUSP and NUCLE corpora and divided into three sub-corpora: the MICUSP-NS (writing by native speakers in MICUSP), the MICUSP-NNS (writing by non-native speakers in MICUSP), and the NUCLE-NNS (writing by a subgroup of non-native speakers in NUCLE). For MICUSP-NNS, all 31 texts written by 31 learners are selected. For the other two sub-corpora, systemic sampling was used to randomly select 31 texts from 155 essays in MICUSP-NS and 31 texts from 1414 essays in NUCLE-NNS.

4.2.4.1. Systemic sampling versus stratified sampling for MICUSP-NS

Since there are 155 native speakers' essays and 31 were to be selected, the systemic sampling rule is to select every fifth essay, starting with a random essay (from the first to the fifth). All possible samples created from this sampling technique are compared to MICUSP-NNS with regard to disciplines and presented in Table 6.

Table 6

Systematic sampling for MICUSP-NS and the number of essays across the disciplines

	MICUSP-NNS	MICUSP-NS				
Disciplines		**#1**	**#2**	**#3**	**#4**	**#5**
BIO	1	1	1			
CEE				1		
ECO					1	
EDU		1	1	1		1
ENG	7	12	11	11	12	12
HIS/CLS	1	3	3	3	3	3
IOE			1	1		
LIN		1		1	1	1
NRE	1	1	2	1	1	1
NUR	1			1	1	1
PHI	4	4	3	2	3	3
POL	1	3	4	3	4	3
PSY	5	2	2	2	2	3
SOC	8	3	3	2	3	3

An alternative technique is using stratified sampling method. To select 31 essays from 155 essays mean to draw one fifth of the population, and hence the proportion of the sample to the population is 1:5, meaning that one fifth of the essays for each discipline should go into the sample. There are, however, disciplines that have a very small number of essays and are not present in the sample (Table 7). Those disciplines are BIO, CEE, ECO, EDU, IOE, LIN, and NUR. A disproportionate stratified sampling to include these minority cases would risk the representativeness of the sample.

Table 7

Stratified sampling for MICUSP-NS

Disciplines	Number of essays by NS in MICUSP	Expected number of essays in MICUSP-NS
BIO	2	0.4
CEE	1	0.2
ECO	1	0.2
EDU	4	0.8
ENG	58	12
HIS/CLS	15	3
IOE	1	0.2
LIN	4	0.8
NRE	6	1
NUR	3	0.6
PHI	16	3
POL	18	3.6
PSY	11	2
SOC	15	5
TOTAL	155	31

In comparison between Table 6 and 7, the systemic sampling is a better option since it loses fewer disciplines than the stratified sampling while still maintains randomness of the selection. According to table 6, sample #1 and #4, created by systematic sampling starting with the fourth essay, show the most similarity to MICUSP-NNS because they both lose one discipline that are present in MICUSP-NNS . Sample #4, however, retains a proportion of essays similar to those in the sample in table 7,

especially across four disciplines PHI, POL, PSY, and SOC. Thus, sample #4 was chosen to be the best representative of native speakers' argumentative writing in MICUSP and to increase the comparability between MICUSP-NS and MICUSP-NNS. Table 8 and 9 presents the sample data in the MICUSP sub-corpora.

Table 8

Argumentative Essays in MICUSP-NS (N_1=31)

Paper ID	Title	Discipline
ECO.G0.03.1	Economics of the Illicit-Drug Market	Economics
ENG.G0.02.1	The Vicar of Wakefield as a Failed Morality Story	English
ENG.G0.06.2	A (Solitary) Place For Fantasy in Reality	English
ENG.G0.12.1	Women in Beowulf	English
ENG.G0.19.2	Paper on Invisible Man for an American Lit course	English
ENG.G0.24.1	Contradiction and Religious Critique: The Pardoner in The Canterbury Tales	English
ENG.G0.29.1	Good People Breaking Rules	English
ENG.G0.37.1	The Last Paper I Ever Wrote in College	English
ENG.G0.42.2	Sexuality in Ancient Greece	English
ENG.G0.49.1	The Purgatory of the Postmodern	English
ENG.G0.53.1	Carwin and the Imp of the Perverse	English
ENG.G1.04.1	Sports Literacy and Rhetoric as Power	English
ENG.G2.03.1	Domesticity in Cold War Black Fiction on the Left	English
CLS.G0.06.1	Analysis of the Parthenon Frieze	History & Classical Studies
CLS.G3.01.1	"Corpse Demons" in Ancient Greek Magic	History & Classical Studies
HIS.G1.03.1	Sex Education in East and West Germany	History & Classical Studies
LIN.G0.12.1	National Identity and Language Education Policy	Linguistics
NRE.G0.11.1	Materials and the Environment	Natural Resources & Environment
NUR.G0.15.1	Circumcision: Challenging a Social Norm	Nursing
PHI.G0.06.6	Emotivism and Solomon's Theory of Emotions	Philosophy
PHI.G0.14.1	Early Fetuses and Constitutional Rights	Philosophy
PHI.G3.01.1	That's Some Fancy Thinkin': Accounting for Pretended Ideas in Hume's Treatise	Philosophy
POL.G0.04.1	Democratization of the European Union	Political Science

POL.G0.12.1	The Role of Party Identification in Voting Behavior	Political Science
POL.G0.41.1	Evgenia Ginzburg's Into the Whirlwind	Political Science
POL.G3.01.1	Measuring Racial Prejudice	Political Science
PSY.G0.23.1	Evaluation of Psychology class	Psychology
PSY.G3.03.1	Culture, Mental Disorders, and Evolutionary Analyses	Psychology
SOC.G0.05.2	The Economics of Poverty	Sociology
SOC.G1.01.2	Reconsidering the Black-White Binary: Where Do We Go from Here?	Sociology
SOC.G3.09.1	Rethinking Marx: Rethinking Race	Sociology

Table 9

Argumentative Essays in MICUSP-NNS(N_2=31)

Paper ID	Title	Discipline
BIO.G0.15.1	Invading the Territory of Invasives: The Dangers of Biotic Disturbance	Biology
ENG.G0.18.1	Individuality and Isolation in Moll Flanders	English
ENG.G0.18.2	Frames and Resistance in Pride and Prejudice	English
ENG.G0.18.3	Satire and Morality in the Vicar of Wakefield	English
ENG.G0.18.4	The Space of Dreams in The Age of Innocence	English
ENG.G0.47.1	Female Bonding in the Novel "Roxana"	English
ENG.G0.58.1	My Reading of Chaucer	English
ENG.G1.06.1	Intergenerational Trauma in Nora Okja Keller's Comfort Woman	English
HIS.G1.04.1	Outsourcing History: On the Necessity of Stepping Out of the Archive	History & Classical Studies
NRE.G1.31.1	Policy decision memo	Natural Resources & Environment
NUR.G0.03.1	What Women Want: Comprehensive Contraceptive Coverage	Nursing
NUR.G0.04.1	Socially-Constructed Stress on Women's Health	Nursing
NUR.G1.07.1	Value in Health Care	Nursing
PHI.G0.15.1	Transaction Ethics	Philosophy
PHI.G0.16.1	Basic Frameworks of Moral Justification and Hierarchy of Human Needs	Philosophy
PHI.G1.02.1	A Defense of Ontological Relativity	Philosophy
PHI.G1.04.1	Explanation and Understanding	Philosophy
POL.G3.02.1	Effect of Anxiety to Citizens' Political Capability	Political Science
PSY.G0.15.1	The Cultural Effects on the Self	Psychology

PSY.G0.34.2	The Trouble with Evan	Psychology
PSY.G1.11.4	Treating psychopathology in adults	Psychology
PSY.G2.10.1	Can projective assessment instruments be helpful with the five common reasons for psychological testing?	Psychology
PSY.G2.10.2	Understanding Dorris's The Broken Cord from a Multiple Systems Perspective	Psychology
SOC.G1.10.1	Interpreting the Theoretical Origin of the Utilitarian, Liberalism and Marxism: Part 1	Sociology
SOC.G1.10.2	Interpreting the Social Theories on Power and Social Knowledge: Part 1	Sociology
SOC.G1.10.6	Interpreting the Theoretical Origin of the Utilitarian, Liberalism and Marxism: Part 2	Sociology
SOC.G1.10.7	Interpreting the Social Theories on Power and Social Knowledge: Part 2	Sociology
SOC.G3.01.2	Rethinking the 'Cultural turn' in Class Formation Theory - To Be a Marxist or Not?	Sociology
SOC.G3.01.3	Rethinking Power in Marx	Sociology
SOC.G3.03.1	The Modern State: Imagined or Real?	Sociology
SOC.G3.06.1	Revolution: Never Obsolete	Sociology

4.2.4.2. Systematic sampling for NUCLE-NNS

Given the large size of the corpus (1414 essays in total) and the essays having not been classified in terms of topic, demographic information and so on, systematic sampling is more suitable than other random sampling methods. Since 1414/31 = 45.6, thirty one texts were systematically sampled by selecting every forty-fifth essay, and the starting point was picked randomly from 1 to 45. In this thesis, 3 was chosen as the starting point. The writers' main arguments of the essays are presented in Table 10.

Table 10

Argumentative Essays in NUCLE-NNS (N_3=31)

Paper ID	Writers' main arguments
"A3"<DOC nid="895">	Physical prototypes and potential profitability affect technological design of 3D printers.
"A48"<DOC nid="1037">	RFID (surveillance tech) should not be banned because it increases public safety and improves identification of patients.
"A768"<DOC nid="2574">	RFID (surveillance tech) should not be banned.
"A1263"<DOC nid="980">	The use of surveillance technology such as RFID to

	track people should be banned
"A1353"<DOC nid="2623">	Surveillance technology should be banned.
"A138"<DOC nid="1122"> "A228"<DOC nid="1270"> "A363"<DOC nid="1374"> "A408"<DOC nid="1184"> "A633"<DOC nid="1090"> "A723"<DOC nid="1296">	Government should cut spending on the aged.
"A678"<DOC nid="1709"> "A273"<DOC nid="1396"> "A1038"<DOC nid="1096">	Public spending on the elderly should not be reduced.
"A588"<DOC nid="949">	Patients should not be the only decision maker to disclose their genetic testing results.
"A813"<DOC nid="2516">	Patients should disclose their genetic testing results.
"A498"<DOC nid="2397">	VHTR (Very High Temperature Reactor) should be supported.
"A903"<DOC nid="2185">	There should be cautions in deciding to implement SCWR (a nuclear power design).
"A948"<DOC nid="2275">	VHTR (Very High Temperature Reactor) should be supported.
"A1083"<DOC nid="2168">	Generation IV reactors should be support.
"A1308"<DOC nid="2305">	Nuclear power should be supported.
"A1218"<DOC nid="2280">	SFW and SCWR (nuclear power designs) should be supported
"A93"<DOC nid="913">	Engineering design process can help solve future energy shortage. (?)
"A183"<DOC nid="1626">	Both well-researched and serendipitously discovered technologiesb face problems. (?)
"A318"<DOC nid="1664">	Engineers' audacious imagination could provide tremendous breakthroughs. (?)
"A453"<DOC nid="1606">	Iphone's remodeling is driven by users' demand for functionality and their financial problems in the economic crisis
"A543"<DOC nid="1498">	Water need should be given the priority in engineering design.
"A858"<DOC nid="1780">	Technology relieves the consequences of the aging process.
"A993"<DOC nid="2428">	Hybrid cars should be improved for safety reasons regardless of high cost.
"A1128"<DOC nid="2153">	Psychologists can help prevent wrong engineering decision making. (?)
"A1173"<DOC nid="1530">	Engineering design process needs adaptation to reality. (?)

Note: The question marks "(?)" indicate, from my own viewpoint, the development of the essays inadequately clarify the thesis statement, as compared to that of the others.

4.2.4.3. Comparability among the three corpora

Based on the features of the MICUSP-NNS, MICUSP-NS, and NUCLE-NNS, the following table has been drawn up to describe the comparability between the three corpora:

Table 11

Comparison of parameters of MICUSP-NS, MICUSP-NNS, and NUCLE-NNS

Parameter	*MICUSP-NS (1)*	*MICUSP-NNS (2)*	*NUCLE-NNS (3)*	*Comparability*	
				(1) and (2)	*(1), (2), and (3)*
Student levels	University final-year undergraduates and up to third-year graduates	Final undergraduates and up to third-year graduates	University undergraduates	FAIR	FAIR
Size (word count)	98340	71343	20121	LOW	LOW
Mean length per essay (word count)	3172.26 (min=566, max=9596)	2317.52 (min=1166, max=6261)	649.06 (min=353, max = 1233)	LOW	LOW
Median length of essay	2623	1903	658	LOW	LOW
Medium	Written	Written	Written	HIGH	HIGH
Text type	Argumentative	Argumentative	Argumentative	HIGH	HIGH
Function	Persuasive, argumentative, literary analysis	Persuasive, argumentative, literary analysis	Argumentative	FAIR	FAIR
Genre	Student's essay	Student's essay	Student's essay	HIGH	HIGH
Audience level	University professors in academic courses	University professors in academic courses	University insructors in academic writing courses	HIGH	FAIR
Discipline family	Arts and Humanities, Life Science, and Social Sciences	Arts and Humanities, Life Science, and Social Sciences	Academically driven with topics in Life Sciences and Social Sciences	FAIR	FAIR

Year of compilation	2006	2006	2009	HIGH	HIGH

4.3. The AntConc Concordancer 3.3.5w

AntConc is a freeware concordance program developed by Laurence Anthony, Waseda University (Japan). Like other commercial programs such as WordSmith Tools and MonoConc Pro, AntConc provides functions to analyze language phenomena and highlight language aspects of interest in a corpus. Römer and Wulff (2010) praised the usefulness of AntConc, asserting that "Without a concordance program like AntConc, a corpus would be of no use other than being an electronic repository of texts that could then be read on screen (or on paper printouts) in the normal linear fashion." It creates a simple way to examine the data by placing the selected items (key words) in the middle of the screen with some context left and right of it. Researchers can access the full text by clicking on the highlighted key words. The concordance also provides graphic plots to depict the position of selected items within the text. Other functions, including creating a word list and keyword list, is irrelevant to the current study.

Chapter 5: Data Analysis and Results

5.1. Data retrieval, categorization, and transformation

5.1.1. Data retrieval

The samples were not immediately available for the analysis. In the NUCLE corpus, the essays are documented in one SGML (Standard Generalized Markup Language) file format with the error tags and corrections. Each essay has no clear identification for researchers to single out. The error tags and corrections, despite its usefulness, may entangle the research results because if researchers use MS Words or AntConc to detect how often *and* appears in the main texts, they may receive results of *and* in the error tags as well. In MICUSP, the essays are stored in an online interactive interface: viewers can select the options as to what types of essays they want to examine, and as they do so the data and the graphs are continuously updated. Users, however, cannot manipulate the data in any way they wish due to limited options available on the website.

Therefore, I manually extracted 93 essays for the three sub-corpora from the NUCLE and MICUSP after having assigned all 1601 essays with an ID number (1414 for NUCLE and 186 for MICUSP) to facilitate the analysis, and stored them separately in both TXT. and DOCX. formats. Between the two formats, only the TXT can be recognized by the AntConc. Figure 1 shows how TXT texts appear in the AntConc and how it orderly structures all the words *and* (called *hits* of *and*). Users can click on the *and* in highlight to access the whole text containing it.

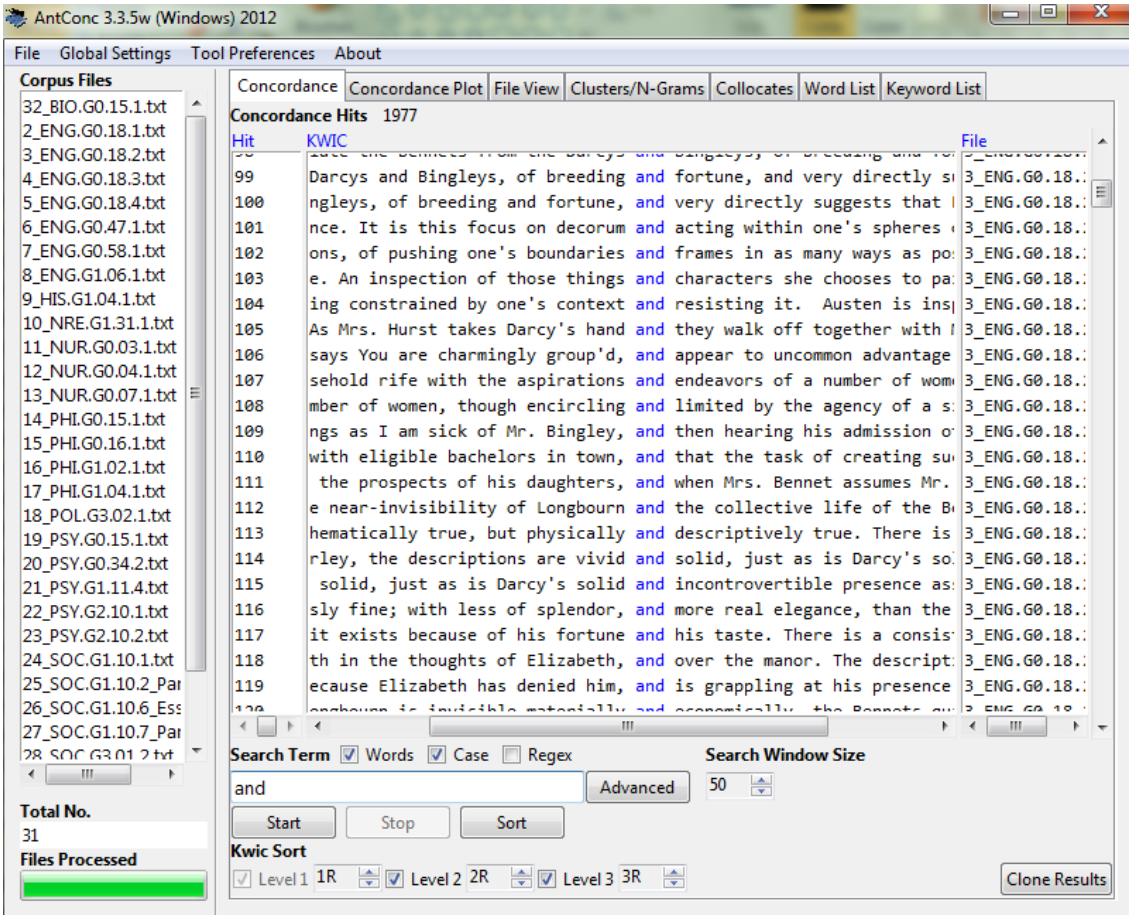


Figure 1. An example of texts and hits of *and* displayed in AntConc.

5.1.2. Data Categorization

To extract instances, or hits, of *and*-conjunctions, the word *and* was put into the search box with the case-sensitive selection on so all *and*-initials (being capitalized) would be eliminated. The results were then exported into a txt.file for printouts. Case by case of *and* was reviewed on the printouts and onscreen, and encoded into SPSS files (SPSS is a software package for statistical analysis). *And* in in-text citations and quotes were also excluded from the results and the analysis.

The categorization of *and*-coordinators followed as closely as possible the description in table 3, chapter 3. In reality, *and* coordinates grammatical structures in many directions, so various that make the categorization in table 3 less feasible. The task

required a heuristic approach in which the researcher had to make adjustments based on new data. The examples below typify some of my decisions.

a. …including increased ***profit*** of export-oriented agri-businesses ***and*** other special ***treatments*** of Thai export in the U.S. market (MICUSP-NNS).

→ *And* connects two noun phrases in which the head nouns are *profit* and *treatment*. The constituents that modify the head nouns include adjective and prepositional phrases. Although there are an uneven number of words filling in each modifying slots, i.e. *increased* versus *other special*, and *export-oriented agri-businesses* versus *Thai* export, the sentence is classified as syntactic symmetry because it has the same-level, main constituents of a noun phrase: [Adj phrase] [Head Noun] [Prep Phrase]. The mismatch between *profit* (head noun 1 in singular form and without any article) and *treatments (*head noun 2 in plural form) are tolerated and acceptable.

b. **These research communities *are closely tied*** to external actors through… ***and*** consequently **their research agenda *represents*** the interest of trans-national companies (MICUSP-NNS).

→*And* coordinates two clauses. The comparison of the components between the two clauses reveals the differences: the head verb in the first clause is in passive voice and followed by a prepositional phrase, while the second head verb is in active voice and followed by a noun phrase. The whole sentence, thus, is considered syntactically asymmetric.

c. …the safety **and/or** protective measures would be more stringent …

→ The conjuncts, *safety* versus *protective,* belong to two distinct part of speech, i.e. noun and verb. The sentence, consequently, is considered syntactically asymmetric.

d. In these ways, the power of masculine sports rhetoric ***has regulated***, and ***continues to regulate***, sports in such a way that…

→ *And* connects two verb phrases but they are unequal because the first verb is in present perfect tense, hence accompanied by the auxiliary verb *has* while the second verb is in simple present tense and accompanied by an ordinary verb (to-infinitive). The conjunction structure of the sentence, as a result, is asymmetry.

5.1.3. Data transformation

The essays are of unequal length (as shown in table 11), so the longer the essay, the more *and*-coordinator it has. The number of *and* in this case is commonly called *the observed absolute frequencies,* and it is impractical to compare this value across essays. Instead, the frequency of *and* in each essay underwent the same normalization procedure to achieve what Gries (2010) calls *observed relative frequencies,* which are typically reported as frequencies per 1,000 or 1,000,000 words. In this research, absolute frequencies of *and* were transformed into relative frequencies per 10,000 by the following formulas:

And-coodinator (within essay) = $\frac{(observed\ absolute\ frequencies\ of\ AND\ in\ the\ essay)*10{,}000}{number\ of\ words\ in\ the\ same\ essay}$

And-coordinator (whole corpus) = $\frac{(total\ absolute\ frequencies\ of\ AND\ in\ the\ corpus)*10{,}000}{number\ of\ words\ in\ the\ same\ corpus}$

According to the formulas, frequencies of *and*-coordinator in the three corpora and for each essay were translated into relative frequencies, as shown in Table 12, 13, 14, and 15.

Table 12

Total observed absolute and relative frequencies of and-coordinator in MICUSP-NS, MICUSP-NNS, and NUCLE (with the exclusion of quotations and in-text citations)

Corpus	Observed absolute	Observed relative

MICUSP-NS	2758	8115.016
MICUSP-NNS	1931	8602.781
NUCLE	455	7044.076

Table 13

Observed relative frequencies of and-coordinator in MICUSP-NS

Paper ID	Word Count	Observed symmetry	Observed asymmetry
ECO.G0.03.1	3986	130.457	67.737
ENG.G0.02.1	1980	131.313	75.758
ENG.G0.06.2	903	143.965	66.445
ENG.G0.12.1	2059	189.412	43.711
ENG.G0.19.2	1765	237.960	124.646
ENG.G0.24.1	2876	198.192	142.559
ENG.G0.29.1	1588	100.756	113.350
ENG.G0.37.1	2486	213.194	72.405
ENG.G0.42.2	800	187.500	25.000
ENG.G0.49.1	1608	292.289	87.065
ENG.G0.53.1	1559	153.945	160.359
ENG.G1.04.1	5765	343.452	157.849
ENG.G2.03.1	9171	245.339	93.774
CLS.G0.06.1	1654	72.551	84.643
CLS.G3.01.1	5651	148.646	76.093
HIS.G1.03.1	9596	239.683	127.136
LIN.G0.12.1	3455	251.809	118.669
NRE.G0.11.1	1444	180.055	131.579
NUR.G0.15.1	4586	141.736	98.125
PHI.G0.06.6	566	88.339	35.336
PHI.G0.14.1	2384	83.893	75.503
PHI.G3.01.1	8196	91.508	71.986
POL.G0.04.1	2785	175.943	71.813
POL.G0.12.1	2907	134.159	61.920
POL.G0.41.1	3256	113.636	89.066
POL.G3.01.1	2160	166.667	55.556
PSY.G0.23.1	2623	171.559	114.373
PSY.G3.03.1	3277	146.475	149.527
SOC.G0.05.2	2648	83.082	56.647
SOC.G1.01.2	3322	285.972	120.409
SOC.G3.09.1	1284	124.611	77.882

Table 14

Observed relative frequencies of and-coordinator in MICUSP-NNS

Paper ID	Word Count	Observed symmetry	Observed asymmetry
BIO.G0.15.1	1278	125.196	20.790
ENG.G0.18.1	1755	176.638	56.180
ENG.G0.18.2	1745	217.765	125.673
ENG.G0.18.3	1742	223.881	77.640
ENG.G0.18.4	1686	177.936	43.290
ENG.G0.47.1	1618	142.151	89.888
ENG.G0.58.1	2425	218.557	28.860
ENG.G1.06.1	4235	127.509	35.524
HIS.G1.04.1	4247	197.787	49.751
NRE.G1.31.1	3769	198.992	47.170
NUR.G0.03.1	2255	141.907	64.516
NUR.G0.04.1	2652	207.391	92.838
NUR.G1.07.1	2732	219.619	44.248
PHI.G0.15.1	1166	94.340	56.657
PHI.G0.16.1	2106	156.695	126.939
PHI.G1.02.1	6261	84.651	38.314
PHI.G1.04.1	2601	126.874	27.894
POL.G3.02.1	3260	171.779	96.552
PSY.G0.15.1	1903	99.842	38.810
PSY.G0.34.2	1645	133.739	118.519
PSY.G1.11.4	1835	190.736	48.701
PSY.G2.10.1	1767	141.483	45.593
PSY.G2.10.2	2316	198.618	97.324
SOC.G1.10.1	2612	222.052	171.184
SOC.G1.10.2	2025	325.926	41.667
SOC.G1.10.6	1199	200.167	24.390
SOC.G1.10.7	1664	204.327	138.462
SOC.G3.01.2	1687	177.830	47.037
SOC.G3.01.3	1395	258.065	118.203
SOC.G3.03.1	2941	153.009	41.783
SOC.G3.06.1	1321	174.111	62.208

Table 15

Observed relative frequencies of and-coordinator in NUCLE-NNS

Paper ID	Word Count	Observed symmetry	Observed asymmetry
"A3"<DOC nid="895">	481	83.160	20.790
"A48"<DOC nid="1037">	356	224.719	56.180
"A93"<DOC nid="913">	557	143.627	125.673
"A138"<DOC nid="1122">	644	341.615	77.640
"A183"<DOC nid="1626">	693	115.440	43.290

"A228"<DOC nid="1270">	445	404.494	89.888
"A273"<DOC nid="1396">	693	245.310	28.860
"A318"<DOC nid="1664">	563	195.382	35.524
"A363"<DOC nid="1374">	402	74.627	49.751
"A408"<DOC nid="1184">	636	157.233	47.170
"A453"<DOC nid="1606">	465	215.054	64.516
"A498"<DOC nid="2397">	754	251.989	92.838
"A543"<DOC nid="1498">	452	44.248	44.248
"A588"<DOC nid="949">	353	28.329	56.657
"A633"<DOC nid="1090">	709	141.044	126.939
"A678"<DOC nid="1709">	783	76.628	38.314
"A723"<DOC nid="1296">	717	97.629	27.894
"A768"<DOC nid="2574">	725	96.552	96.552
"A813"<DOC nid="2516">	773	77.620	38.810
"A858"<DOC nid="1780">	675	266.667	118.519
"A903"<DOC nid="2185">	616	178.571	48.701
"A948"<DOC nid="2275">	658	167.173	45.593
"A993"<DOC nid="2428">	1233	113.544	97.324
"A1038"<DOC nid="1096">	701	156.919	171.184
"A1083"<DOC nid="2168">	720	138.889	41.667
"A1128"<DOC nid="2153">	820	219.512	24.390
"A1173"<DOC nid="1530">	650	76.923	138.462
"A1218"<DOC nid="2280">	1063	131.703	47.037
“A1263"<DOC nid="980">	423	141.844	118.203
"A1308"<DOC nid="2305">	718	181.058	41.783
"A1353"<DOC nid="2623">	643	139.969	62.208

5.2. Results from descriptive statistics

In descriptive statistics, the three populations, represented by the three samples MICUSP NS, MICUSP-NNS, and NUCLE-NNS, can be described by a set of parameters, which includes the measure of central tendency (or description of the central location of the sample data by such descriptors as mean and median) and the measure of variability (description of the amount of spread of the sample data by the standard deviation and range).

Descriptive statistics showed, non-native speakers in MICUSP tended to use more *and*-coordinators in total than the other two groups (M_2=277.51 versus M_1=261.77 and

M_3=227.23), and non-native speakers in NUCLE used *and* the least (M_3=227.23) (Table 16). The medians approximated the means, indicating the means to be relatively a reliable parameter of the populations from which the samples was drawn and, therefore, can be used in inferential parametric statistics. The standard deviation of *and* frequencies in MICUSP-NNS showed that the frequencies of *and* in MICUSP-NNS were more clustered around the mean than the other two groups (SD_2=64.255 versus SD_1=89.458 and SD_3= 97.036), which means the MICUSP-NNS sample data are more normally distributed than data in MICUSP-NS and NUCLE-NNS or, in other words, more a representative of its advanced non-native speakers population. The range for NUCLE-NNS is the highest (409.40), indicating potential outliers (essays with extremely high or low frequencies of *and* as compared to the remaining) that may skew the distribution of the frequency variable and, hence, reduced its normality. The boxplots in Figure 2 singled out observation 68 (“A228”<DOC nid=”1270”>) as the outlier of NUCLE-NNS.

Table 16

*Summary statistics of **observed relative frequencies** of total and-coordinators in MICUSP-NS, MICUSP-NNS, and NUCLE*

Parameters	**MICUSP-NS (N_1=31)**	**MICUSP-NNS (N_2=31)**	**NUCLE-NNS (N_3=31)**
Mean (M)	261.77	277.51	227.23
Median	233.12	278.59	215.38
Standard Deviation (SD)	89.458	64.255	97.036
Range	377.63	279.494	409.40
Minimum-Maximum	123.68 – 501.30	150.14 – 429.63	84.99 – 494.38

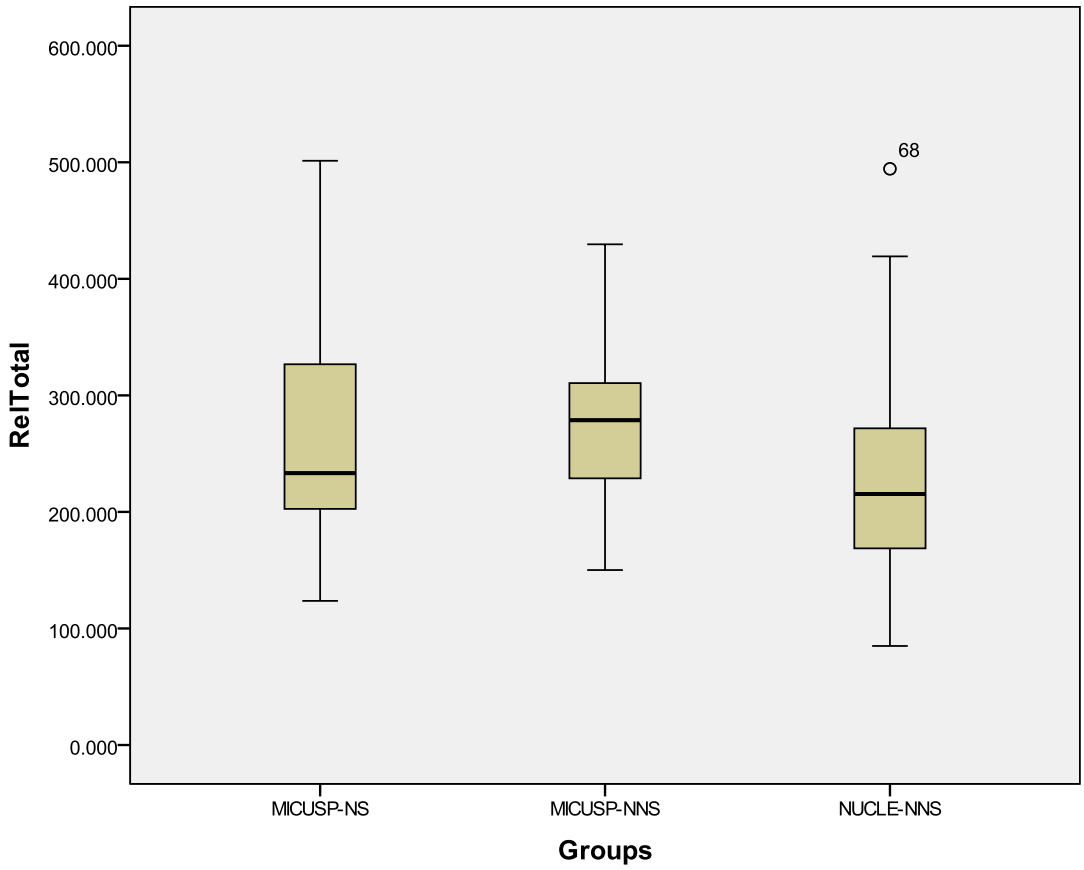


Figure 2. Boxplot for the summary statistics in Table 16.

In Table 17, the essays in MICUSP-NNS demonstrated the highest average use of *and* in their symmetric structures (M_2=177.08), and those in NUCLE-NNS the least (M_3=158.95). The means of the three groups approximated the medians; thus, the mean is a good indicator of the population mean and appropriate for parametric tests. The median for MICUSP-NS signifies half of the essays contained fewer than 153.94 *and* in syntactically symmetric structures, and the other half have more than 153.94. Similar interpretations of the median can be applied to the medians for MICUSP-NNS (177.83) and NUCLE (141.84).

The standard deviation of *and* frequency in its syntactic structure was the least for MICUSP-NNS (SD=50.83), as compared to MICUSP-NS (SD=67.883) and NUCLE-

NNS (SD=84.053), which indicates the data in MICUSP-NNS was more clustered around the mean, while data in NUCLE-NNS were the most dispersed. This scattering distribution of NUCLE-NNS can be partly explained by the highest range (R=376.17) among the three groups.

Table 17

*Summary statistics of **observed relative frequencies** of and-coordinator in syntactically symmetric structures in MICUSP-NS, MICUSP-NNS, and NUCLE*

Parameters	**MICUSP-NS (N_1=31)**	**MICUSP-NNS (N_2=31)**	**NUCLE-NNS (N_3=31)**
Mean (M)	169.94	177.08	158.95
Median	153.94	177.83	141.84
Standard Deviation (SD)	67.883	50.83	84.053
Range (R)	270.90	241.27	376.17
Minimum-Maximum	72.55 – 343.45	84.65 – 325.93	28.33 – 404.49

The summary statistics were presented in boxplots and detected three outliers: observations 12 (ENG.G1.04.1), 56 (SOC.G1.10.2), and 68 (“A228”<DOC nid=”1270”>).

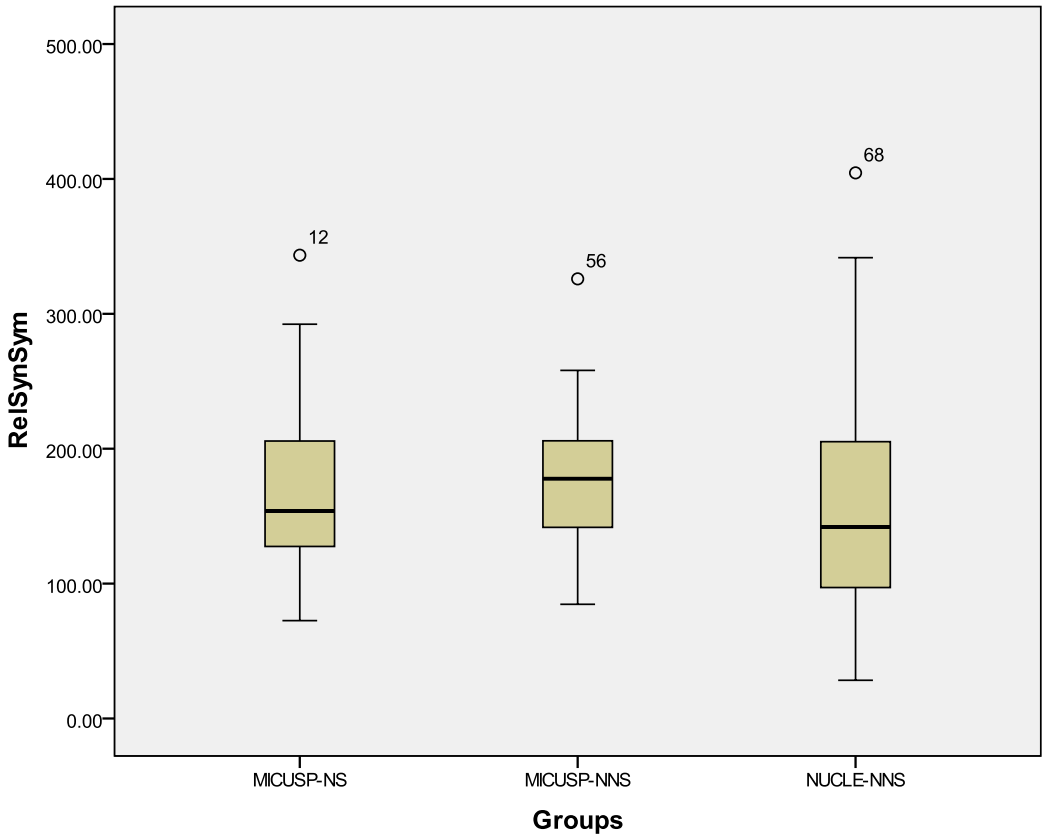


Figure 3. Boxplot for the summary statistics in Table 17.

With regard to and-coordinator in syntactically asymmetric structures, the lowest average use of *and* was observed in NUCLE-SNSS (*M*=68.28), while higher average occurrences of *and* were found in MICUSP-NS (*M*=91.94) and MICUSP-NNS (*M*=100.43) (Table 18). For MICUSP-NS and MICUSP-NNS, the median was close to the mean, which establishes the mean as a trustworthy parameter. For NUCLE, the median and the mean were more apart. With the standard deviation somewhat high (SD_3=38.883), as compared to MICUSP-NS (SD_1=35.720) and MICUSP-NNS (SD_2=29.240), and as compared to NUCLE-NNS mean and median, the NUCLE-NNS median seems to be another parameter that should be considered in inferential statistics.

Consistent with its standard deviation, NUCLE-NNS showed the most dispersion of data, ranging from 20.79 (the lowest value) to 171.18 (the highest value).

Table 18

*Summary statistics of **observed relative frequencies** of and-coordinator in syntactically asymmetric structures in MICUSP-NS, MICUSP-NNS, and NUCLE*

Parameters	**MICUSP-NS (N_1=31)**	**MICUSP-NNS (N_2=31)**	**NUCLE-NNS (N_3=31)**
Mean (M)	91.94	100.43	68.28
Median	84.64	96.21	49.75
Standard Deviation (SD)	35.720	29.240	38.883
Range (R)	135.36	116.17	150.39
Minimum-Maximum	25.00 – 160.36	42.29 – 158.47	20.79 – 171.18

Figure 4 illustrates the statistics in Table 18 and shows there were no outliers in all three corpora. Outliers in Figure 2 and 3 were not eliminated during the data analysis because the Kolmogorov-Smirnov's test was not violated in cases of total *and* and *and*-symmetry (Table 19). In other words, the outliers did not make distributions significantly depart from normality. Moreover, there were no outliers in the case of *and*-asymmetry.

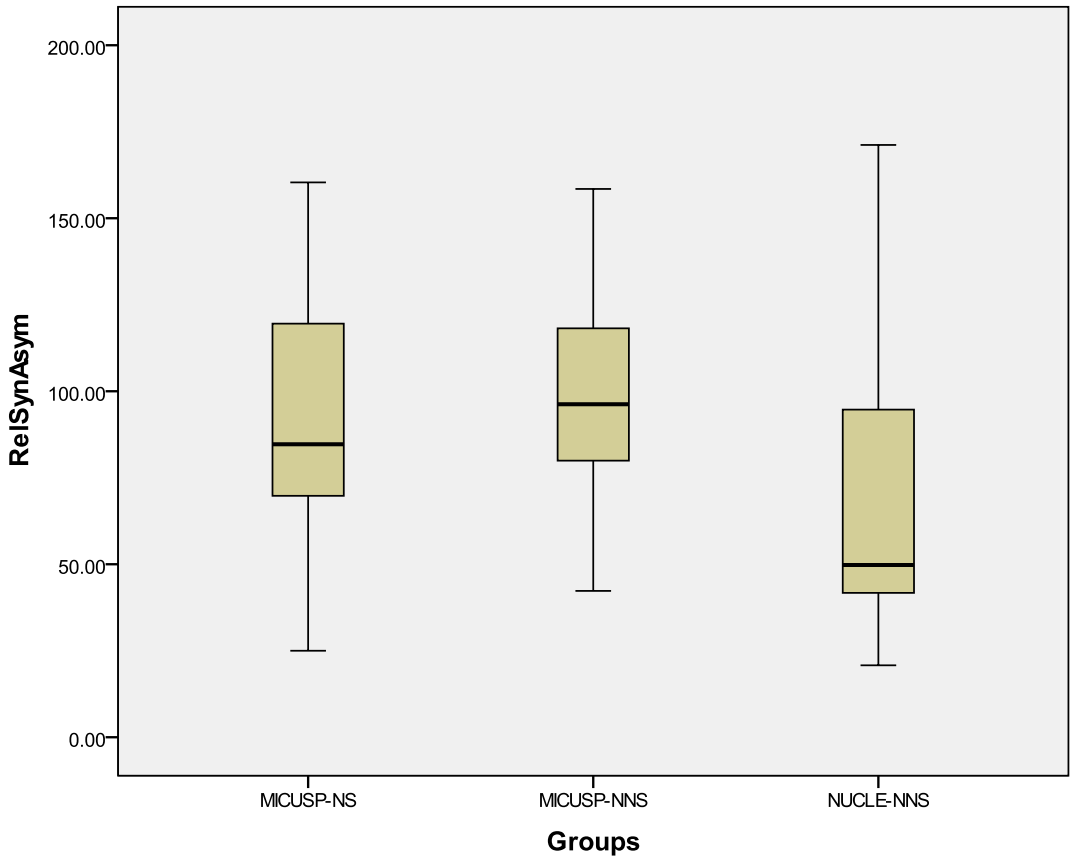


Figure 4. Boxplots for the summary statistics in Table 18.

In general, descriptive statistics suggests that advanced non-native speakers in MICUSP-NNS group tended to use *and*-coordinator more often in total and in two syntactic categories, advanced native speakers in MICUSP-NS used less *and*-coordinator, and upper-intermediate non-native learners in NUCLE-NNS used *and*-coordinator the least.

5.3. Results from inferential statistics

To make further claims about the populations represented by the three samples, inferential statistics was computed on the data set.

5.3.1. Parametric test: One-way ANOVA

Parametric tests allow researchers to make inferences about the population based on the sample mean, assuming the sample mean is fairly similar to the population mean, hence a reliable indicator of the population. In this research, the one-way between subjects analysis of variance (ANOVA) was conducted to compare means frequency of *and*-coordinator in total and in syntactically (a)symmetric structures for the three independent groups MICUSP-NS, MICUSP-NNS, and NUCLE-NNS.

5.3.1.1. Assumptions for one-way ANOVA

The one-way ANOVA assumes independent samples, homogeneity of variances, and normality of distributions for the samples. The completely randomized design of this research (sampling method) ascertained the samples were independent. The Levene's test qualified assumptions for homogeneity of variances for three learners' groups with regard to total frequency of *and*-coordinator (p=.118), frequency of *and*-coordinator in syntactically symmetric structures (p=.085), and frequency of *and*-coordinator in syntactically asymmetric structures (p=.174). The Kolmogorov-Smirnov's test of normality of distributions confirmed normality for total frequency of *and*-coordinator and frequency of *and*-coordinator in syntactically symmetric structures across three corpora, but not for frequency of *and*-coordinator in syntactically asymmetric structures in NUCLE-NNS (Table 19).

Table 19

Results of Kolmogorov-Smirnov's test of distribution normality

	MICUSP-NS	MICUSP-NNS	NUCLE-NNS	OVERALL THREE CORPORA
Total frequency of *and*-coordinator (for hypothesis 1)	p=.095	p=.200	p=.200	p=.200
Frequency of *and*-coordinator in syntactically asymmetric (for	p=.200	p=.200	p=.200	p=.200

hypothesis 2)				
Frequency of *and*-coordinator in syntactically asymmetric structures (hypothesis 3)	p=.153	p=.200	p=.003	p=.200

Because the assumptions for hypotheses 1 and 2 are qualified, one-way ANOVA would yield highly accurate results. Further discussion about ANOVA for hypothesis 3 will be presented in the following sessions.

5.3.1.2. One-way ANOVA test results for the hypotheses

One-way ANOVA test predicted no significant difference in means frequency of *and-coordination* in argumentative academic writings among the three learner groups MICUSP-NS (M=261.77, SD=89.458), MICUSP-NNS (M=277.51, SD=64.255), and NUCLE-NNS (M=227.23, SD=97.036), $F(2, 90)=2.855$, p=.063.

The test also showed no significant difference in means frequency of *and-coordination* in syntactically symmetric structures in argumentative academic writings among the three learner groups MICUSP-NS (M=169.94, SD=67.883), MICUSP-NNS (M=177.08, SD=177.83), and NUCLE-NNS (M=158.95, SD=84.053), $F(2, 90)=.544$, p=.582

For the third hypothesis, although Kolmogorov-Smirnov's test suggested a non-normal distribution of frequency of *and-coordination* in syntactically asymmetric structures in the population for NUCLE-NNS group (Figure 2), the non-normal distribution has little impact on the accuracy of the ANOVA test because a) the sample size of NUCLE-NNS is large enough (N=31) that, according to the central limit theorem, the sample means should distribute normally or, in order words, the sample mean would approach the population mean; b) there were no outliers in terms of asymmetric

coordination in any of the three group MICUSP-NS, MICUSP-NNS, and NUCLE-NNS, and; c) the skewness (light-tailed) of the data for NUCLE-NNS group was less than 1.

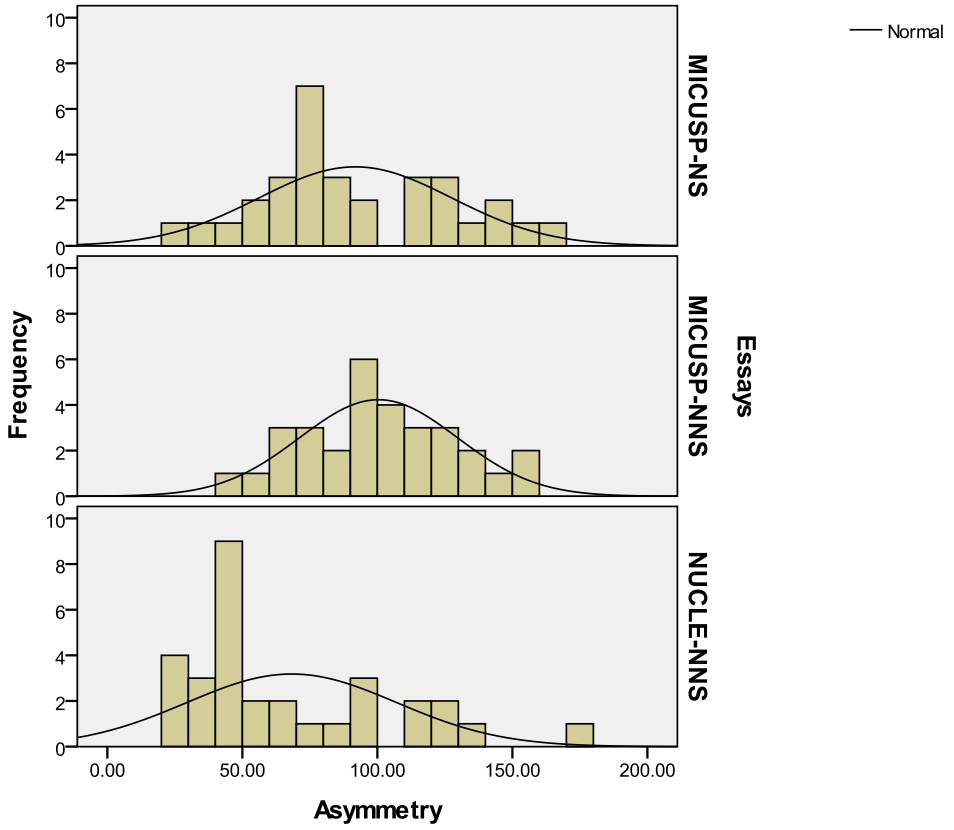


Figure 5. Distributions of relative frequencies of *and*-coordinator in syntactically asymmetric structures in the three corpora MICUSP-NS, MICUSP-NNS, and NUCLE-NNS. The normal curve superimposing the histograms represents the normal distribution of data predicted by the central limit theorem.

One-way ANOVA test showed that the effect of language proficiency on frequency of *and* regarding its syntactically asymmetric coordination in argumentative academic writings was significant, $F(2, 90)=7.073$, $p=.001$, $\eta^2= .136$. Tukey's post hoc procedure indicated, learners in NUCLE-NNS group ($M=68.28$, $SD=38.883$) used *and*-asymmetric coordination significantly less than those in MICUSP-NS group ($M=91.94$, $SD=35.720$) and in MICUSP-NNS group ($M=100.43$, $SD=29.240$). There was not a

significant difference in the number of *and* in asymmetric structures between the two groups MICUSP-NS and MICUSP-NNS.

Given the controversy that linguistic data are non-normally distributed, the Kruskal-Wallis test, a non-parametric test independent of normality assumption and equivalent to ANOVA was applied to the data analysis.

5.3.2. Non-parametric test: Kruskal-Wallis test

The Kruskal-Wallis test uses the median, instead of the mean, as the parameter for comparison. The test requires assumption of similarity of distributions. According to Figure 2, the normality curves look fairly similar; thus, the shape of the distributions of the populations can be assumed to be similar.

The Kruskal-Wallis test was conducted to evaluate the differences among three different learner groups (MICUSP-NS, MICUSP-NNS, NUCLE-NNS) on median change in frequencies of *and* in asymmetric structures. The test was significant $X2(2, N = 93) = 13.766$, $p = .001$. The proportion of variability in frequencies of *and* in asymmetric structures was accounted for by proficiency level was 14.96 %, indicating a strong relationship between proficiency level and frequency of *and* in asymmetric structures.

Follow-up Mann-Whitney U-tests were conducted to evaluate pair-wise differences among the three learner groups. Results indicate a significant difference between the NUCLE-NNS group and the MICUSP-NS group, $p = .008$. The frequency of *and* in asymmetric structures was higher for MICUSP-NS than for NUCLE-NNS. Results also indicate a significant difference between the NUCLE-NNS group and the MICUSP-NNS group, $p = .000$. The frequency of *and* in asymmetric structures was higher for

MICUSP-NNS than for NUCLE-NNS. Frequency of *and* in asymmetric structures did not differ significantly between MICUSP-NNS and MICUSP-NS, $p = .263$.

In summary, both one-way ANOVA and Kruskal-Wallis tests confirmed significant relationship between proficiency level in academic writings and the frequency of *and*-coordinators in asymmetric structures. Other variables that might have an impact on the outcome will be discussed in chapter 6.

Chapter 6: Discussion of Results, Contributions, and Limitations

6.1. Discussion of findings

6.1.1. Explanations for the significant difference in frequency of syntactically asymmetric and-coordination across high levels of proficiency

The quantitative results suggested the upper-intermediate college learners may use much fewer *and*-asymmetric structures than higher level students, but it requires a close-up of *and* instances to account for the difference. The examination of *and*-asymmetric structures in four categories—conjoined sentences, NPs, VPs, and APs—across the three corpora revealed that, advanced learners use more *and-asymmetry* in academic writings may be because they tend to expand or compress complex ideas within one sentence. Thus, they use various grammatical structures, such as relative clauses, infinitive clauses, prepositional phrases, and so on, to create larger constituents and simultaneously maintain cohesion. The larger those constituents are, the more possibly the coordination becomes asymmetric. This point is demonstrated through the following analysis of several examples of *and*-asymmetric structures.

> (1) However, as will be shown, ***the costs associated with these problems*** are tremendous, and (=so) ***we*** must now call into question whether or not current forms of <u>prohibition and resource allocation</u> are an effective and efficient means of attacking these problems.

→ While the first conjunct has a reduced relative clause (or participle phrase), the second one is embedded with a *whether*-clause containing two symmetrical coordination structures (MICUSP-NS).

(2) Another negative externality often associated with illegal drug use is increased incidences of property crime. Because ***many drug offenders*** have no legitimate means of income, **and as was noted earlier, *there* are no positive profits to be made in the long-run at the street level,** offenders often resort to property crime as a means of subsidizing their habits (MICUSP-NS).

→ *And* connect two conjuncts within the subordinate clause (the clause begins with subordinate conjunction *because*). The idea in the second conjunct is a nice reminder of another idea previously mentioned. This makes the whole sentence more cohesively related to the essay.

(3) ***Her faith*** in her Governess here is out of desperation**, and** soon enough, ***she*** is back to her occupation and independence (MICUSP-NNS).

→ The prepositional phrase *in her Governess* specifies *her faith*, and the second sentence contains a symmetrical coordination.

(4) ***Anonymity*** is the vehicle through which Moll is a success at robbery, **and *her preservation of her identity*** keeps her in the business much longer than the rest; she is not a member of the gentry that she robs from, because they must not recognize her, but she also is not a member of the society of thieves (MICUSP-NNS).

→ The asymmetric structure is caused by the embedded clause in the predicate in first conjunct (*through which Moll is a success*), the prepositional phrase in the subject in the second conjunct (*of her identity*), the prepositional phrase in the predicate of the second conjunct (*in the business*) and adverbial phrase in comparative form *(much longer than the rest)*.

(5) The Potsdam palace is still standing. ***Its mass <u>of stone and mortar</u>*** <u>has retained most of its shape and weight</u>, **and *it*** is still furnished with what passes for the best of rococo elegance

→ The first conjunct has two symmetric coordinating structures and the second one a passive voice and a free relative clause, making the sentence asymmetric and complex. The meanings behind the two conjuncts are unanimous in describing the Potsdam palace.

(6) In particular, a major problem that must be called to the attention of ***policy makers,*** **and *the economists who advise on their behalf***, is the <u>observation and recognition</u> of a nearly perfectly inelastic supply of drug offenders-dealers-at the street level (MICUSP-NS).

→ *And* coordinates two conjuncts below the predicate level (i.e. prepositional phrase). The second conjunct includes a subject-relative clause. It shows, with the possessive pronoun *their* referring to and in close proximity to *policy makers',* a strong cohesive relationship with the first conjunct.

(7) It is a philosophy of freedom, of choice, of ***letting go of that which is unimportant*** **and *accepting a world characterized by <u>irrationality and indifference</u>*** (MICUSP-NS).

→ The two conjoined NPs have gerunds as the heads, a relative clause as a modifier of object-preposition *that*, and a reduced relative clause (*characterized by)* with a symmetric coordination (*irrationality and indifference)*

(8) In Pride and Prejudice as well as in Evelina, an important distinction is necessitated between ***those who understand*** **and *practice etiquette and those who are truly polite and good-natured*** (MICUSP-NS).

→ The third *and* conjoins two NPs within a PrepP, each contains a relative clause differing with regards to the predicate level.

(9) Given consumers' trends of GMOs rejection in the world market including the US, Thailand could instead be more competitive by ***remaining a non-GM food producer*** **and** ***focusing on the organic food market*** (MICUSP-NNS)**.**

→ *Remain* functions as a copular verb while *focus* does not. But the reason this coordination is asymmetric is the prepositional phrase following *focusing*.

(10) Since the advent of reproductive control methods, women have struggled to find a balance between ***what they need*** **and** ***what they can get*** (MICUSP-NNS)**.**

→ The asymmetric degree in this case is not much because the two conjuncts are almost syntactically similar. What make this syntactically complex is the free relative clause (Wh-clause) functioning as a NP within a PrepP.

(11) Some of the behavioral symptoms that he displays that supports the presence of this disorder are ***his blunted emotion, increase in aggressiveness,*** **and** ***avoidance of places associated with the stimuli*** (MICUSP-NNS)**.**

→ The conjuncts show a variety in syntactic structures, with the head nouns modified by participle, PrepP, and PrepP with reduced relative clause respectively.

(12) Accordingly, it cannot be ignored that HIV, an extremely costly disease to diagnose and treat***, is epidemic among prison populations*** **and, hence,** ***only worsens the monetary crisis now being felt by correction institutions*** (MICUSP-NS)**.**

→ The asymmetry is characterized by the copular verb *be* conjoined with the transitive verb *worsen*, and the participle phrase modifying the noun *crisis.*

(13) While Plunkett' s textual identity is ***governed by*** **and** ***constitutive of*** a discourse of British imperialism, however, he is nonetheless compelled to construct a familial genealogy in a mode of discourse more mythical than historical (MICUSP-NS).

→ This asymmetric coordination involves connecting verb in passive voice with an adjective.

(14) Instead, she ***further harms her relationship with her family*** **and** ***creates distrust in her ability to be a mother*** (MICUSP-NNS)**.**

→ The infinitive-clause *to be a mother* breaks the symmetry of the coordination structure. In this sentence, the writer can use *motherly ability* instead of *ability to be a mother*.

(15) Since having a cavalry was ***a new*** **and** ***probably quite an exciting*** occurrence for the city of Athens, it would make sense that a procession by this powerful military force would be quite exhilarating for <u>those watching, and even those taking part in it</u> (MICUSP-NS).

→ The adjective phrase in the second conjunct comprises of an adjective and adverbs while the first one only has one adjective. This asymmetry emphasizes the writer's comment implied in *probably* and *quite.*

(16) Though it is likely that more and more countries will be forced to welcome GM produce, the ***safety*** **and/or** ***protective*** measures would be more stringent as more obvious segregation of GMOs and non-GMOs (MICUSP-NNS).

→ *And* connects the noun phrase *safety* and the adjective phrase *protective*, making the coordination a simple asymmetric structure.

Advanced learners even intentionally employ *and*-asymmetry for stylistic choices.

(17) ***The actual linguistic behavior of the foreigner physicist*** yields no difference in favor of one or the other; and ***no other possible linguistic behavior*** will (MICUSP-NS).

→Ellipsis occurs in the second sentence, which helps avoid unnecessary repetition and establishes the syntax asymmetry.

(18) During the attack, she ***proves her physical superiority over the men*** **and** ***manages to take two worthy prizes*****:** the king's favorite man and the arm of her son (MICUSP-NS).

→The asymmetry is characterized by an unequal component between the two conjuncts: the noun phrase *her physical superiority* following the head verb *proves* and the infinitive-clause *to take two worth prizes* following the head verb *manages.* The asymmetry in ***the king's favorite man and the arm of her son*** seems to be intentional in order to create rhythmic balance.

(19) For example, we may perceive someone's enthusiasm for their culture as over-bearing or even a little inappropriate, when we are really neglecting the fact that they may have been raised to ***be very knowledgeable about*** **and** ***have great pride for*** their heritage (MICUSP-NNS).

→ The conjuncts are asymmetric because the first conjunct is constituted by copular *be* and adjective phrase while the second one by an ordinary verb and noun phrase. The meaning in the second conjunct seems to receive more emphasis than the first.

(20) Only then would a movement come about that could **challenge** the feudal elite sufficiently to achieve land reforms, **and** in that final formal act, **do away with** them all together (MICUSP-NNS).

→ The asymmetry is created within the relative clause. *And* connects the two unbalanced verb phrases: the first one is followed a noun phrase with an infinitive clause, and the second one followed by a noun phrase and an adverb (*together).* The asymmetry in this sentence is associated with stylistic choice: the writer used comma before the second conjunct to create a short break, and, with the second conjunct shorter than the first in length, to emphasize the act *do away with.*

In contrast, upper-intermediate college learners who enroll in academic writing courses may not have such a profound, discipline-oriented knowledge that advanced learners have already acquired. Their lack of knowledge about the academic topic hinders the attempt to combine grammatical constituents to expand their ideas. In cases when upper-intermediate learners develop or compress their ideas and, thus, create *and*-asymmetric coordination, the asymmetric coordination shows more redundancy, vagueness, and less cohesion. *And*, therefore, is relegated to its very basic sense: connecting two equal elements, but without much attention paid to the semantics and stylistic choice underlying each.

(21) Moreover, with a higher level of education***, they*** will be able to learn how to invest well, which will eventually lead to the more salary they will be able to earn **and *it*** will result in the increase in the Gross Domestic Product, which is good for their country (NUCLE-NNS).

→ The conjuncts are sentences, each modified by a relative clause, and for the first relative clause another relative clause is embedded within it. The relative clauses, *which will eventually lead to the more salary they will be able to earn* and *which is good for their country*, however, express consequences based on common sense rather on academic (or text-based) evidence, as compared to the examples in MICUSP-NS and MICUSP-NNS. The referential *it* is vague as readers are not certain whether *it* refers to *the more* salary or to the idea of good investments. The cohesive relationship of meanings between the conjuncts is not clear enough. Therefore, the conjuncts, although asymmetrical and complex, seem wordy and redundant.

(22) The by-products of VHTR are hard to extract and the usable amount is of insignificant amount (NUCLE-NNS).

→ The asymmetry is not a reflection of high level proficiency. In fact, it emerges because of the writer's lack of vocabulary: the writer was not sure which adjective has equivalent meaning to *insignificant amount.*

(23) On the one hand, nuclear energy has the advantage of ***high efficiency, none-greenhouse-gas emission*** and ***the by-product hydrogen which is supposed to be another ideal energy*** (NUCLE-NNS)**.**

→ The third conjunct (*by-product hydrogen*) needs an adjective before it because in the other two conjuncts, the word *high* and *none* create readers' expectation of quantity clarifying each noun in the coordination structures.

(24) Even though it may be costly as extensive research has to be carried out and expensive construction is required ***(S)***, the benefits and advantages ***(N)*** will definitely outweigh this hefty price. Importantly, global issues such as

greenhouse effect and depleting non-renewable resources (N) <u>are being addressed and can be solved (V)</u> (NUCLE-NNS)<u>.</u>

→ The conjuncts *greenhouse effect* and *depleting non-renewable resources* differs in the part-of-speech of the elements modifying the head nouns. They also differ in the inflectional morpheme *–s* for plural nouns (i.e., resources). This difference might have caused the writer to make a mistake with using articles: while plural nouns do not require *the,* a singular countable noun, e.g. *greenhouse effect*, does.

(25) The new one has higher speed core and bigger RAM, which lets Iphone runs faster and can even display High definition movie (NUCLE-NNS).

→ The writer had difficulties with verbs such as *let* that take bare-infinitive. Thus, the use of *and* in this asymmetric structure causes confusion as to what noun phrase, RAM or Iphone, is the subject of *can even display.*

(26) Thus, elderly ***will feel convenient to do their daily activities*** **and** ***live independently*** (NUCLE-NNS).

→ The coordination is asymmetric because of the modal verb and the infinitive clause in the first conjunct, which do not occur in the second conjunct. The idea in the second conjunct seems to be a repetition of the first one.

(27) Given the structure of the present hybrid cars, there are indeed some areas of safety aspects that are neglected and should be included to increase the safety of users (NUCLE-NNS).

→ The asymmetry in this structure involves modal verb and is quite syntactically simpler than those in MICUSP examples.

(28) Hence, spending on pension and retraining for the elders ***is essential*** **and** ***should not be limited*** as it aids them in being financially independent after retirement (NUCLE-NNS).

→ The asymmetry in this sentence is syntactically simple and not due to stylistic choice. In fact, the first conjunct can be omitted completely.

(29) From this train accident, factors begin from the preliminary wheel design that design engineers came up with, to ***the insufficient maintenance of trains*** **and** ***the operation mode ICE uses*** (NUCLE-NNS)**.**

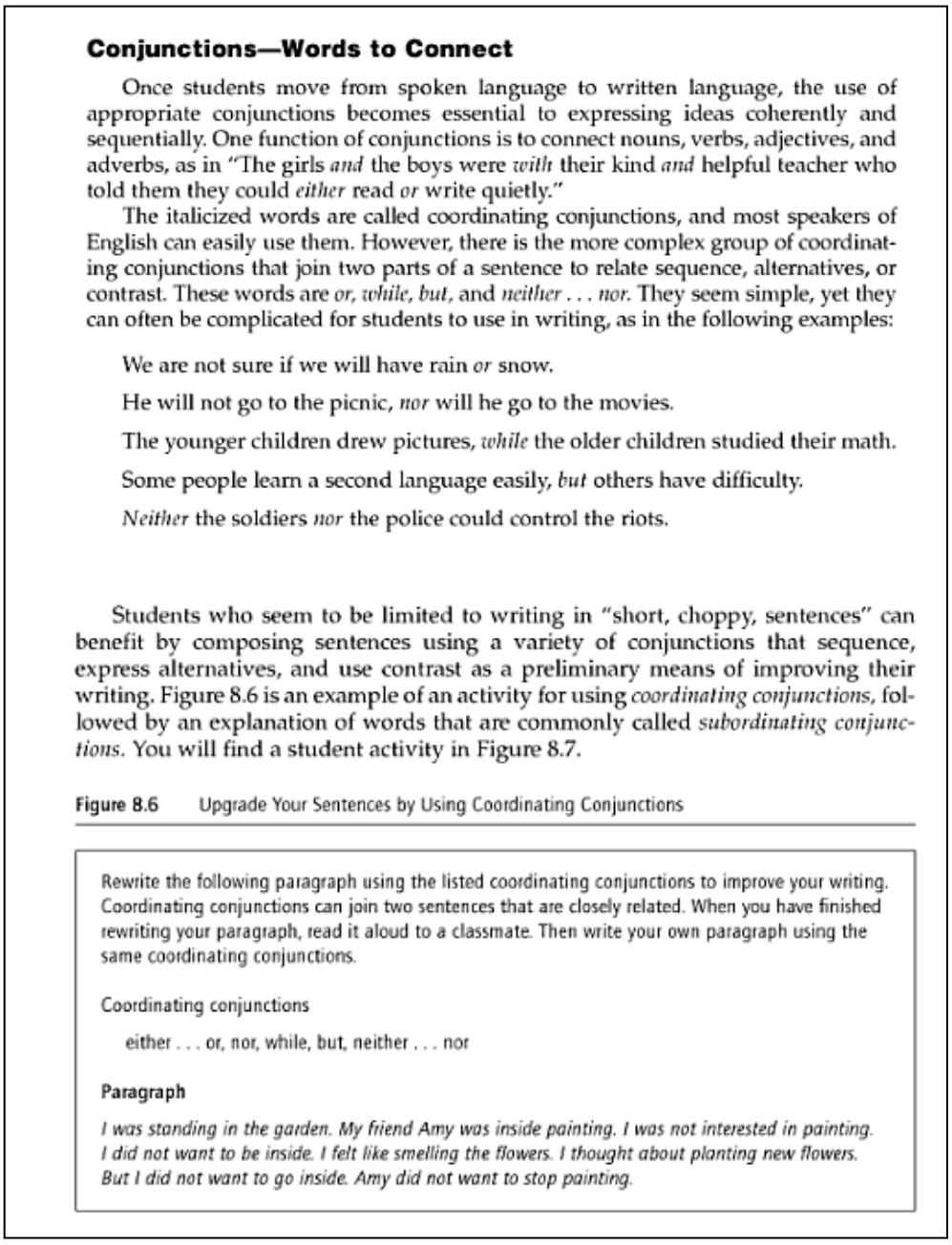
→ Different from other examples, the conjuncts demonstrate a good use of *and* to connect the noun phrase (*train* and *the operation mode ICE uses)*. The second conjunct contains a relative clause (*ICE uses*). This is not a common case in the NUCLE-NNS corpus.

What can be loosely generalized from my observation of the NUCLE-NNS corpus is English learners starting college and at upper-intermediate proficiency may have not been much exposed to or internalized intricate grammatical structures, such as complex sentences joined by subordination, coordination and relative clauses. In writing, they may resort to simple structures they acquired at lower-levels, such as sentences with simple S+V+O and no embedded clauses and phrases consisting minimally of a head. As simple structures are coordinated by *and*, the possibility they are syntactically symmetric increases.

Learners may also avoid *and*-coordination in favor of other grammatical structures, due to their increasing awareness of them or to instructions in academic

writing classes. Instead of saying, "…creating models with traditional means is time consuming, ***and*** does not make economical sense to businesses," a learner wrote, "…creating models with traditional means is time consuming, ***which*** does not make economical sense to businesses" (NUCLE). Not all avoidance of *and* leads to mistakes, but it may reduce the amount of *and*-asymmetric coordination.

Additionally, the emphasis of other conjunctions overtime may shape how students perceive and use the word *and*. Grammar books for students often create an impression that the *and*-coordination is the simplest conjunction. In E. Rothstein and A. Rothstein's grammar instructions for teaching PreK-12 English learners, *and* is mentioned briefly in a short example in which the conjoined structures are symmetrical (Figure 6). The authors explicitly exclude *and* from the list of what they call "complex group of coordinating conjunctions" while other coordinators, such as *but* and *or,* are on the list (p.132). For ESL's classroom, logical connectors (e.g., *before*, *besides* and *moreover*) are taught from an early age, and language teachers often encourage using them in writing (Leung, 2005). Leung ascribes the overuse of connectors and the underuse of *and*-coordinators to such emphasis in teaching English. Leung's argument implies instructions from teachers on using conjunctions may be influential enough to sway learners' language behavior even when learners can incorporate both logical connectors and *and*-coordinators in the same sentence.



Conjunctions—Words to Connect

Once students move from spoken language to written language, the use of appropriate conjunctions becomes essential to expressing ideas coherently and sequentially. One function of conjunctions is to connect nouns, verbs, adjectives, and adverbs, as in "The girls *and* the boys were *with* their kind *and* helpful teacher who told them they could *either* read *or* write quietly."

The italicized words are called coordinating conjunctions, and most speakers of English can easily use them. However, there is the more complex group of coordinating conjunctions that join two parts of a sentence to relate sequence, alternatives, or contrast. These words are *or, while, but,* and *neither . . . nor.* They seem simple, yet they can often be complicated for students to use in writing, as in the following examples:

We are not sure if we will have rain *or* snow.

He will not go to the picnic, *nor* will he go to the movies.

The younger children drew pictures, *while* the older children studied their math.

Some people learn a second language easily, *but* others have difficulty.

Neither the soldiers *nor* the police could control the riots.

Students who seem to be limited to writing in "short, choppy, sentences" can benefit by composing sentences using a variety of conjunctions that sequence, express alternatives, and use contrast as a preliminary means of improving their writing. Figure 8.6 is an example of an activity for using *coordinating conjunctions,* followed by an explanation of words that are commonly called *subordinating conjunctions.* You will find a student activity in Figure 8.7.

Figure 8.6 Upgrade Your Sentences by Using Coordinating Conjunctions

Rewrite the following paragraph using the listed coordinating conjunctions to improve your writing. Coordinating conjunctions can join two sentences that are closely related. When you have finished rewriting your paragraph, read it aloud to a classmate. Then write your own paragraph using the same coordinating conjunctions.

Coordinating conjunctions

either . . . or, nor, while, but, neither . . . nor

Paragraph

I was standing in the garden. My friend Amy was inside painting. I was not interested in painting. I did not want to be inside. I felt like smelling the flowers. I thought about planting new flowers. But I did not want to go inside. Amy did not want to stop painting.

Figure 6. An excerpt about coordinating conjunctions from E. Rothstein and A. Rothstein's grammar book.

Because English learners bring in their own L1 background to study the target language, how they use *and*-coordinators may also be influenced by L1, and this phenomenon is referred to as *forward transfer*—a form of language transfer. Learners in

MICUSP-NNS and NUCLE-NNS do not share the same L1, but the majority of them speak Chinese as their mother tongue. Therefore, the research results might reflect a transfer substantially from Chinese to English. The findings that upper-intermediate learners use *and*-asymmetric structures significantly less than native speakers supports Leung's claim that language transfer is in effect (2005). If forward transfer occurs, why do advanced non-native learners not use *and*-asymmetry at the same rate?

Chen's study of Chinese university learners' L2 writing (1999) points out the interaction of L2 proficiency with L1 transfer: L1 transfer is influential at low levels of L2 but dissipates when learners become more competent of L2. Although Chen's research dedicates to another grammatical structure, his conclusion might extend to the current study as well based on the fact that the upper-intermediate group (i.e., NUCLE-NNS) uses *and*-syntactical asymmetry much less than the advanced non-native group (i.e., MICUSP-NNS). The diminution of L1 transfer could be accounted for by the "ignorance hypothesis," which maintains that learners might recourse to their L1 habit to substitute for their lack of L2 knowledge (Jarvis & Pavlenko, 2008, p.8). In this case, like Leung's comment, they could simply carry over their L1 habit in using *and* because they are not familiar with L2 usage patterns. The use of *and*-asymmetric coordination, thus, becomes a case of L1 transfer linearly correlating with proficiency in L2 while in many other circumstances, language transfer "does not decrease linearly as competence and proficiency in [L2] increase" (Jarvis & Pavlenko, 2008, p.11).

In short, whether it is the mentality of treating *and* as a simple conjunction or language transfer that comes into play, these factors seem all connect to learners' L2

proficiency. Figure 7 summarizes possible explanations for the differences in the use of *and*-asymmetric structures among upper-intermediate and advanced learners:

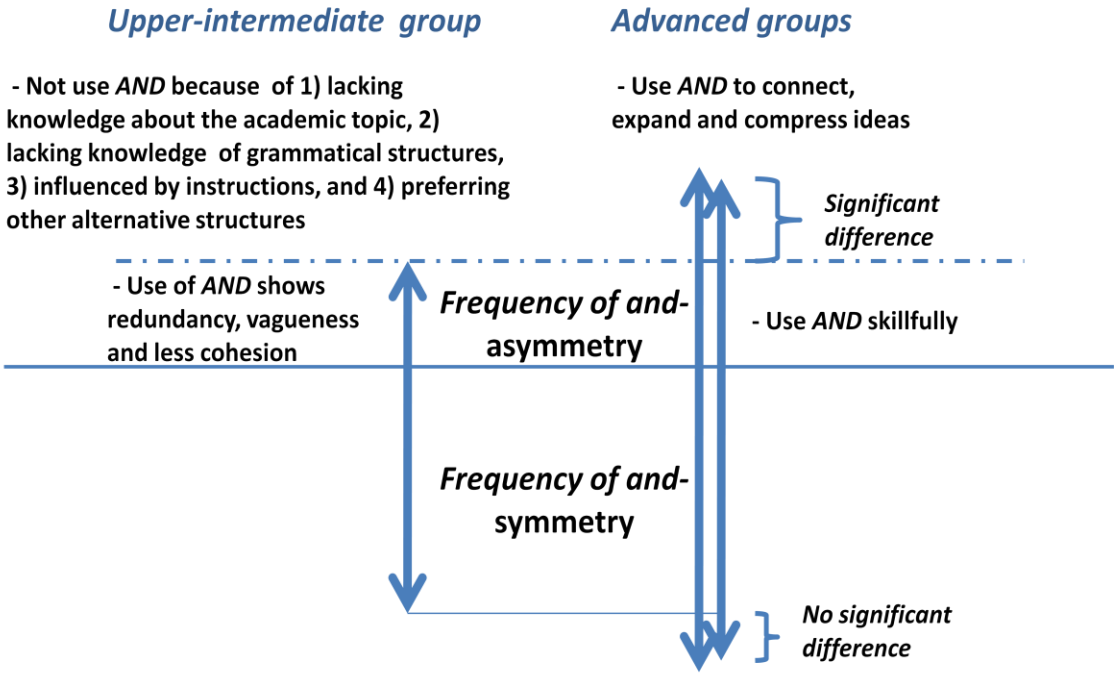


*Figure 7.*Comparison of *and*-syntactical asymmetry among three learner groups: upper-intermediate non-native speakers, advanced non-native speakers, and advanced native speakers.

6.1.2. *Elaboration on the non-significant difference in frequency of syntactically*

(a)symmetric *and*-coordination across high levels of proficiency

According to the non-statistically significant results, there is not enough evidence that advanced non-native speakers use *and* differently from advanced native speakers in both asymmetric and symmetrical categories in academic argumentative writing. The result alludes to the possibility that when their proficiency is as nearly high as that of native speakers, the proficiency becomes less of a decisive factor for the difference in frequency, but it does not mean proficiency has no influence. The result rather proposes a necessity to incorporate other factors, such as stylistic choice, topic, and discipline, rather

than proficiency, to compare between writing by non-native and native speakers, and channels attention back to the qualitative approach. Consider the following example:

> "Although most situations where two informed parties agree to a contract, ***un-coerced*** **and** ***better off*** afterwards without producing harm to anyone else, are permissible, there are cases for which these seemingly exhaustive principles still result in an ethically questionable transaction" (MICUSP-NNS).

The seemingly benign *and*-syntactical symmetry connecting two adjectives, *un-coerced* and *better off*, within the reduced relative clause, in fact causes readers' confusion: what could be the subject for *un-coerced* and *better off*? Although the writer intended *a contract* to be the subject of *un-coerced* and *better off,* the adjective *better off* often collocates with human subjects, and, being within close proximity to *two informed parties*, creates the impression that it modifies *parties,* hence the confusing *and*-coordination. Therefore, subsequent qualitative research on writing by highly proficient ELL writers may need substantial qualitative analysis before any quantitative study is conducted to categorize potential causes of differences between their writing and native speakers'.

Between the upper-intermediate and advanced groups, the non-significant difference in frequency of *and*-symmetric structure indicates proficiency less affect the use of the structure, but there is no clear evidence to explain this phenomenon. In my follow-up survey of fifteen essays in the MICUSP-NNS, more than two-third of symmetric structures are noun-phrase coordination, following by adjective-phrase, verb-phrase, and sentence. Meanwhile, only about one third of asymmetric structures involve noun phrases as conjuncts, but the proportions of verb phrases-as-conjuncts and

sentences-as-conjuncts are relatively high, much higher as compared to their proportion in symmetry structures; for example, approximately 1.4% of total symmetric versus 23% of total asymmetric structures derived from verb phrases-as-conjuncts. If similar results can be found in the other two corpora, a possible conclusion seems to be that phrasal structures, especially noun phrases, are easier to coordinate symmetrically and, thus, learners across proficiencies are more likely to use them.

6.1.3. Total and-coordination: What other variables may influence the frequency of and besides language proficiency?

Statistics showed no significant differences among the three groups—upper-intermediate non-native, advanced non-native, and advanced native speakers—regarding the frequency of *and*-coordination even though advanced non-native speakers in MICUSP-NNS slightly use more *and* than the other two groups, and upper-intermediate non-native speakers in NUCLE-NNS seem to use *and* the least. The results indicated there was not enough evidence to support the hypothesis that language proficiency may influence the frequency of total *and*-coordination in academic argumentative essays. Further studies on similar issues may need samples with larger sizes in order for researchers to support the hypothesis.

On the other hand, the results are open to questions: whether there are other variables coming into play, and what else could be the differences in use of *and*-coordination among the three groups. As Nesi (2011) demonstrates how the discipline the essay is written for can influence certain features in writing, I grouped the essays in MICUSP-NNS and MICUSP-NS according to general disciplinary groups (Arts & Humanities, Life Sciences, and Social Sciences). Each essay has its specific discipline in

its file name (e.g. ENG as English). I did not group them into topics because they are widely diverse. Meanwhile, I assorted NUCLE-NNS essays into groups of similar topics (Table 10, chapter 4). Although NUCLE-NNS essays were not written for a specific discipline, their topics and contents are roughly equivalent to those in Life Sciences and Social Sciences. For example, the topic *public spending on older adults* may be found in politics, and *nuclear power* in natural resources and environment.

I ran Concordance Plot, a function in AntConc, to generate visual images of instances of *and*-coordinator in the form of a barcode for each essay (See Appendix). Each vertical line in the barcode represents the position of the word within the text (Römer & Wulff, 2010). The texts used for the concordance plot are those including quotes and in-text citations, so the visual plot also counts those *and*-coordinators learners do not actually produce, but the number of those cases is relatively small, so it does not affect much the visual plot. Interpreting the plot, however, should be done with caution because barcodes are normalized to the same length in visual display, but actual lengths of the essays (in characters or words) are diverse. Thus, the actual width between the two *and*-coordinators (represented by the two vertical lines) is scaled differently for each barcode depending on the actual length (in characters). One can observe in general the longer the essay is, the closer the proximity between the vertical lines in the plot seems to be. Despite this tricky issue, a comparison among barcodes with a consideration of the essay's length can reveal areas that are worth exploring in future research.

The plot displays a few differences across disciplines. The frequency of *and* mirrors its densest distribution in barcode 12 (English), 13 (English) and 16 (history). In contrast, *and* seems much more scattered and much less heavy in classical study (CIS)

and philosophy writing in both MICUSP-NNS and MICUSP-NS. The densest for philosophy is barcode 16 (MICUSP-NNS) and 22 (MICUSP-NS), both having length around 40,000 characters; at approximately this length, the barcodes for English are much more packed, such as barcode 12 (MICUSP-NS, ENG, 37333 characters), and the other shorter ones are trending toward thickness, such as barcode 19 (MICUSP-NS, NUR, 30289 characters), 10 (MICUSP-NNS, NRE, 24875 characters), 24 (MICUSP-NNS, SOC, 16140 characters), and 30 (MICUSP-NS, SOC, 21693 characters). In general, Life Sciences display the steadiest distribution as compared to Arts & Humanities and Social Sciences. There seems to be little influence of topics on *and* dispersion because the dispersion patterns vary within the same topics (NUCLE).

6.2. Theoretical, empirical, and methodological contributions

6.2.1. Theoretical contributions to the definition and classification of and-coordination (grammar), the use of and-coordination in learner language (second language acquisition), and teaching English as a second language in academic writing

The theoretical contribution in this thesis is meaningful, not in the sense of discovering a groundbreaking way to define coordination *and* in the mix of linguistic debates, nor sorting through a dazzling array of linguistic frameworks to propose a better way of classifying *and* with respect to its (a)symmetric features. What the thesis promotes is to reconfigure the common perception and treatment of *and*-coordinators in corpus research. Linguistic theories have constructed an intricate picture of *and;* even its definition is food for thought. Unexpectedly, the majority of corpus studies on *and* are

limited to the general notion of *and* as a signifier of *additive* cohesion at both intra- and inter-sentential levels. What exactly did those studies not mention?

They did not reexamine the definition of *and*-coordination to see how it is reflected in learner language. In *The Grammar Book,* Celce-Murcia and Larsen-Freeman define coordination as joining "two constituents of the same type" (1999, p.461). This definition leaves out instances where the two constituents are dissimilar, to a great extent, such as *read slowly and with great excitement* (an adverb phrase coordinated with a prepositional phrase), and to a lesser extent, *be very knowledgeable about* and *have great pride for* (two conjoined verb phrases differing in their syntactic structures). It excludes cases where the conjuncts are controversially non-constituents, as in *Pierre bought and Paul read the book* (Right Node Raising), and where gapping exists, as in *John trim the tree and Stuart the head.* All these cases are not rare in learner language.

In the thesis, such cases are classified as syntactically asymmetric *and*-coordination although linguists find the (a)symmetry notion elusive. The discussion in chapter 3 attempts to capture a glimpse of this notion through multiple perspectives from classical Generative Grammar and Minimalism to Functional Grammar. These views seem to suggest that a symmetric nature of *and*-coordination is inherent in the equally syntactic level of the conjuncts, even when the two parts appears syntactic asymmetry (i.e., the patterns or types of the conjuncts are somewhat different, as mentioned above). Thus, the thesis proposes a modified version of Celce-Murcia and Larsen-Freeman's definition to clarify the fundamental symmetry of *and*-coordination without excluding asymmetric cases. The syntactic symmetry used as a criterion for the major comparison in the thesis should be understood not as the symmetric nature of *and* (i.e., equally

syntactic level of the conjuncts), but as the uniformity of the patterns of the conjuncts in comparison to asymmetric structures.

The rationale behind the categorization is not only linguistically theory-driven. When learner language is the context for the comparison, the decision to adopt the categorization is at first intuitive in that some learners might find asymmetric structures more difficult to produce in writing than symmetric ones. The additional error analysis, the empirical approach, and the close-up inspection of learners' writing support the intuition. The classification, therefore, is justifiable.

Another theoretical contribution is to the field of second language acquisition, namely, how learners beyond the intermediate level of English proficiency use *and*-coordination. The analysis of learners' academic writing indicates a movement of use through developmental stages from upper-intermediate to advanced levels: as learners reach the advanced level, they seem to use the syntactically asymmetric *and*-coordination significantly more than at the lower level (upper-intermediate), and they use it at a similar rate as compared to their advanced native peers.

The finding is open to question whether there is an interaction between language transfer and proficiency level. Studies on forward transfer show that the direction of L1 transfer is not always predictable when L2 proficiency increases; it can decrease, increase, or remain the same (Jarvis & Pavlenko, 2008). Chen (1999) suggests that for Chinese students, L1 transfer tends to decrease when L2 proficiency accumulates. Leung (2005), who found Chinese undergraduate students underuse *and*-coordination in writing for English language courses, suspects L1 transfer influences the results. Since a large portion of the participants in the current thesis are Chinese, the finding might be another

case of language transfer interacting with L2 proficiency. Future research could focus on the syntactically asymmetric *and*-coordination and compare results across different learners' L1 to elucidate the relationship between transfer and L2 proficiency.

The final contribution is to teaching English grammar to speakers of other languages. The traditional teaching model corresponds to learners' levels of English proficiency, and it is based on classification of grammatical structures into levels of difficulty (Meunier, 2002). The current approach is communicative-based, targeting learners' need and putting grammatical structures into contexts of use (Meunier, 2002). The thesis endorses the second approach for teaching *and*-coordination in academic writing context: the use of *and* should be taught not only at low levels, but at all levels of proficiency, and in relation with other grammatical structures and the academic content. This viewpoint will be expanded in section 6.3, "Implication for Teaching ELLs."

6.2.2. Empirical contributions to previous research on the use of conjunction and

The thesis attracts attention to the *and*-coordinator at the phrasal level, from my perspective *an underexplored territory of the Wild West*. The research incorporates inferential statistics to confirm that upper-intermediate English learners use less *and*-asymmetric coordinator at phrasal and sentential levels combined than their advanced peers, while advanced English learners use *and*-coordinator as a similar rate as their advanced native peers. This finding expands on the conclusion that non-native speakers use significantly more coordinating conjunctions than native speakers, as in Reid's study (1992) in which the author examined 768 English essays for comparison/contrast/take-a-position and interpretation-of-a graph topic types written by native and non-native speakers.

The research generates empirical evidence that supports Leung's study (2005), in which Leung stops short at using descriptive statistics. By selecting large samples and employing inferential statistics, the thesis enables a generalization of the finding. The result is backed up with both parametric and non-parametric tests to cover the possibility that linguistic data could be non-normal.

The follow-up examination, including visual data from a dispersion plot, presents how non-native learners use *and* differently from their peers regarding what to combine with *and* and what discipline the essays were written for. From my observation, the use of *and* differs in syntactic complexity, underlying semantics of the conjuncts, and stylistic choices. Errors related to *and* and vocabulary words can happen at even highly advanced levels. *And*-conjunctions distribute evenly in several long essays, and in general the distribution varies widely regardless of the essays' topics or the disciplines they belong to. Visual plots indicate philological essays have the least number of hits, and the hits are more scattered as compared to those in other disciplines. These findings suggest that future research could further explore the use of *and* and other cohesive devices in writing in the disciplines.

This thesis also sets an example of searching for differences in the use of *and* by adjusting the criteria of comparison. Previous studies have focused tremendously on total frequency of *and* as an additive conjunction and its sentence-initial position. This thesis, instead of counting on the frequency of *and* in general, narrows the criteria down to syntactic (a)symmetry of *and.* Its findings imply if researchers look into, for example, semantic (a)symmetry, vocabulary, proximity, and prosody, or even narrow further down to, let's say, embedded clauses, they may discover a wide range of usage patterns that

amount to differences between learner groups. I believe there may be more about *and* than the common conclusion that ELLs (or even native speakers) overuse *and.*

6.2.3. Methodological contributions to previous research in corpus linguistics

Regarding methodological contributions, the thesis addresses several issues related to terms that are often overlooked by previous studies. Confusing terms such as *genres, text types, corpus-based, corpus-driven, deductive approach,* and *inductive approach* are distinguished. This clarification helps characterize any corpus studies and writing data under those studies' investigation.

The non-uniformity in evaluating proficiency levels for learners is another issue that hinders comparison across corpora and corpus studies (Guo, 2006). The research resolves this issue by nominating an intermediary yardstick— the ACTFL Proficiency scales. There are other guidelines that could serve the same purpose, such as the Common European Framework of Reference for Languages, but in the context of the MICUSP— an American corpus, the ACTFL is more suitable.

6.3. Implications for teaching ELLs

Even though upper-intermediate ELLs tend to use much less *and*-asymmetric coordinator in academic writing than their advanced peers, it would be impractical to encourage them to link ideas with more *and* without careful instructions because, as discussed in section 6.1.1, they have difficulties with *and*-asymmetric coordination. They may generate more ineffective *and*-coordination in their essays if the teacher urges them to use *and.* In addition, overusing a specific pattern of *and*-asymmetric structures can make the essay less appealing, as in "I got up early, and I finished studying. I went to school, and I took the algebra quiz…" (Elliot, 2006).

The observation of *and*-asymmetric coordination in writing by native speakers and advanced ELLs offers a solution: English instructors should integrate *and* with subordination, (reduced) relative clause, ellipsis, to-infinitive clause, and other syntactic structures in examples they give to upper-intermediate ELLs. They need to demonstrate such combinations can help express ideas concisely. In order for ELLs to improve *and*-coordination, instructors can ask them to clarify the semantic relationship between the two conjuncts and read more about the academic topic, so they can create *equal importance* in meaning (i.e., not redundancy and irrelevance) of the conjuncts.

An implication derived from *equal importance in meaning* is that practice tasks in which ELLs connect asymmetric conjuncts at various syntactic levels may enhance their critical thinking because they have to understand what main elements are to be connected and how to maintain the parallelism of sequences.

Finally, given undergraduate ELLs often overuse some adverbial conjunctives such as *furthermore* and *therefore* as discussed in chapter 2, instructors may find it beneficial for upper-intermediate ELLs to connect their ideas using *and* instead of adverbial conjunctives, provided the meaning of *and* is clearly restricted by the context, and using *and* allows more flexibility to generate various structures for the conjuncts.

6.4. Limitation of my study

One limitation of my study is the categorization of *and*-asymmetry. It depends on the principle that the symmetrical sequences are those very similar in syntax while asymmetrical conjuncts differ from each other. Thus, cases like "in our life and in our contemporary society" are examples of asymmetric structures. The benefit of the classification, however, is questionable, because such asymmetry does not distinguish

between learners at and beyond upper-intermediate level. A new classification should be based on learners' error analysis and features they struggle with at lower levels.

Another issue is related to the reliability of the categorization since one person did all the counting and sorting *and.* It would be better to have two people categorizing the *and* words, and after that, the homogeneity of the two classifications is calculated in terms of inter-rater reliability.

Finally, the study does not control disciplines the essays belong to. It is not clear how much disciplines contribute to the differences in addition to language proficiency.

Although results from comparing the three corpora are the basics for my discussion on upper-intermediate and advanced level, I am well-aware that strictly speaking, the findings can only be generalized to the populations to which the samples belong, i.e., native and non-native speakers at high levels of college at the University of Michigan, and non-native speakers studying academic writing at the National University of Singapore.

6.5. Suggestions for future research

Subsequent research can attempt to classify syntactic categories of *and* differently until a better account for the difference among groups is found. For example, researchers can limit the categories to NP only and analyze how many PrepP or embedded clauses are used to modify the head. They can also explore how language proficiency influence semantic (a)symmetry in writing. Before the study is conducted, however, researchers may need a preliminary study to refine the categories because the categorization suggested in chapter 3 is too simple.

Another direction is to explore the use of *and* across disciplines. For example, this thesis found that philosophical essays seem to have the least number of *and*, but the reason for this observation has not been found. As discussed in chapter 2, section 2.3., texts that are more logic-oriented tend to use less coordination (Smith & Frawley, 1983). It might be the case that philosophical essays are more logical than other texts. A quick survey of philosophical texts in MICUSP-NNS and MICUSP-NS reveal students use the subordinator *if* and modal verb *will* a lot more than writers in other disciplines, but how this pattern of use is related to the use of *and*, and whether the content or the topic of the text limit writers in using *and* or not is still open to questions. Another observation is that philosophical essays contain subordinators, but their frequencies seem not higher than those in other disciplines. If philosophical texts are more logic-oriented than the others, then this observation might indicate that a great deal of cohesion in philosophical texts are not signaled by conjunctions.

Future research may also need to collect texts written by below-advanced level with longer text length. It means that the text written by one person should be much longer in NUCLE, as text length to a certain extent seems to influence the density of *and*.

Finally, subsequent studies can explore the relationship between parallelism and syntactic asymmetry in the *and*-coordination. A syntactically asymmetric *and*-coordination, as in "…we are really neglecting the fact that they may have been raised to **be very knowledgeable about and have great pride in their heritage**" (MICUSP-NNS), creates a better parallelism than the symmetric *and*-coordination in its alternative equivalence, "raised to **be very knowledgeable about and greatly proud of** their

heritage**.**" An account for why native speakers perceive parallelism as such would contribute to research on *nativeness* in second language studies.

References

Alarcon, J. B, & Morales, K. N. S. (2011). Grammatical cohesion in students' argumentative essay. *Journal of English and Literature, 2*(5), 114-127.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in Swedish learners' written English. In S. Granger (ed.), *Learner English on Computer* (pp. 80-93). New York, NY: Longman.

Argamon, S. and Dodick, J. (2004). Conjunction and modal assessment in genre classification: A corpus-based study of historical and experimental science writing. In *Notes of AAAI Spring Symposium on Attitude and Affect in Text.*

Artstein, R. (2005). Coordination of parts of words. *Lingua, 115,* 359-393.

Barnbrook, G, Viana, V., & Zyngier, S. (Eds.). (2011). Methodological and interdisciplinary stance in Corpus Linguistis. *Perspectives on corpus linguistics* (pp. 81-98). Philadelphia, PA: John Benjamins.

Bawarshi, A. S., & Reiff, M. J. (2010). *Genre: An introduction to history, theory, research, and pedagogy*. West Lafayette, IN: Parlor Press.

Behrens, B. (2008). Explaining advanced L2: Discourse structural properties of coordinating conjunction in English L1 and advanced L2. In L. Degand, C. Fabricius-Hansen, & W. Ramm (Eds.), *Linearisation and Segmentation in Discourse* (pp. 17-30). Oslo: University of Oslo.

Bell, D. (2007). Sentence-initial *and* and *but* in academic writing. *Pragmatics, 17*(2), 183-201.

Ben-Anath, D. (2005). The Role of Connectives in Text Comprehension. *TESOL & Applied Linguistics, 5*(2), 1-27.

Berzlanovich, I, & Redeker, G. (2012). Genre-dependent interaction of coherence and lexical cohesion in written discourse. *Corpus Linguistics and Linguistic Theory, 8*(1), 183-208.

Biber, D. (1988). *Variation across speech and writing*. New York, NY: Cambridge University

Blakemore, D., & Carston, R. (2005). Introduction to coordination: Syntax, semantics and pragmatics. *Lingua, 115,* 353-358.

Blakemore, D., & Carston, R. (2005). The pragmatics of sentential coordination with *and*. *Lingua, 115,* 569-589.

Bolton, K., Nelson, G., & Hung, J. (2002) A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong. *International Journal of Corpus Linguistics, 7(*2), 165-182.

Butler, C. S. (2003). *Structure and function: From clause to discourse and beyond.* Philadelphia, PA: John Benjamins.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston, MA: Heinle.

Chen, C. W. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics, 11*(1), 113-130.

Chen, F. J. (1999). *The roles of first language transfer and second language proficiency in the writing of Chinese learners of English as a second language*. (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia.

Chiu, Y. (2004). The L2 acquisition of the coordinating conjunction *and* in Taiwanese learners interlanguage at college level. *Asian EFL Journal, 6*(3), 1-20.

Conner, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Paper in Linguistics, 17*(3), 301-316.

Conrad, S. (2005). Corpus linguistics and L2 teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 393-409). Mahwah, NJ: Lawrence Erlbaum Associates.

Crewe, W. J. (1990). The illogic of logical connectives. *ELT J, 44*(4), 316-325.

Dahlmeier, D., and Ng, H.T. (2011). Grammatical error correction with alternating structure optimization. In D. Lin (Ed.), *Annual Meeting of the Association for Computational Linguistics: Human language technologies* (pp. 915-923). Stroudsburg: Association for Computational Linguistics.

De Haan, F. (2010). Building a semantic map: Top-down versus bottom-up approaches. *Linguistic Discorvery, 8*(1), 102.

De Haan, F. (2010). Building a semantic map: Top-down versus bottom-up approaches. *Linguistic Discorvery*, *8*(1), 102.

Diller, H. (2001). Genre in Linguistics and Related Discourses. In H. Diller & M. Görlach (Ed.), *Towards a History of English as a History of Genres* (pp. 3-43). Heidelberg: Universitätsverlag.

Elliott, R. (2006). *Painless grammar.* Hauppauge, NY: Barron's.

Field, Y. and Yip, L. M. O. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC, Journal 23*, 15-28.

Field, Y., & Oi, Y. L. M. (1992). A comparison of internal conjunctive cohesion in English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*, 15-28.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393-417.

Frawley, W. J., & Smith, R. N. (1983). Conjunctive cohesion in four English genres. *Text, 3*(4), 347-374.

Gabrielatos, C. & McEnery, T. (2005). Epistemic modality in MA dissertation. *Lingüística y Filología, 61*, 311-331.

Gardner, S. (2008). Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: an illustration through university history and engineering assignments. In E. Steiner & S. Neumann (Eds.). *Proceedings of the 19th European systemic functional linguistics conference and workshop: Data and interpretation in linguistic analysis.* Saarbrucken: Universitat des Saarlandes.

Gilquin, G. (2011, Sep 30). Center for English Corpus Linguistics. Retrieved from http://www.uclouvain.be/en-cecl-icle.html

Granger, S. (2003). Teachers of English to Speakers of Other Languages. *TESOL Quarterly, 37*(3), 538-546.

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.

Gries, S. Th. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches*, (pp 269-291). Frankfurt am Main: Peter Lang.

Guo, X. (2006). *Verbs in the written English of Chinese learners: A corpus-based comparison between non-native speakers and native speakers.* (Unpublished doctoral dissertation). University of Birmingham, UK.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning.* London: Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* New York, NY: Longman.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning, 12*(2), 111-132.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers, 8*(4), 966-974.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York, NY: Taylor & Francis.

Johnson, P. (1992). Cohesion and coherence in compositions in Malay and English. *RELC J., 23*(1): 1-17.

Laybutt, B. E. (2009). *Collocation and textual cohesion: A comparative corpus study between a genre of Written Sports Reports and a large reference corpus*. (Master's Thesis). Retrieved from www.asian-efl-journal.com/Thesis/Thesis-Laybutt.pdf

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated grammatical error detection for language learners*. San Rafael, CA: Morgan and Claypool.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37-72.

Lee, D. Y. W. (2010). What corpora are available? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp107-121). New York, NY: Routledge.

Leech, G, Rayson, P., & Wilson, A. (2001). Grammatical word class variation within the British National Corpus sampler. *Language and Computers, 36,* 295-306.

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes, 11*(3), 267-275. doi: 10.1016/j.jeap.2012.05.003

Leung, C. (2005). *A comparison of the use of major English conjunctions by American and Hong Kong university students (Using the HKUST corpus, HKBU corpus and the ICLE corpus of American English).* (Master's Thesis). Retrieved from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1331671&fileOId=1331672

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33,* 623-636.

Longacre, R. (1996). *The grammar of discourse.* New York, NY: Plenum Press.

Lorenz, G. (1999). *Adjective intensification—learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

Manchon, R.M., & Norris, J.M. (2012). Investigating L2 writing development from multiple perspectives: Issues in theory and research. In R. M. Manchon (Ed.). *L2 writing development: Multiple perspectives.* Boston, NY: Walter de Gruyter.

Martin, J. R. (1997). Analysing genre: Functional parameters. In F. Christie & J.R. Martin (Ed.), *Genre and institutions: Social processes in the workplace and school* (pp. 3-39). London: Cassell.

Martinez, D., & Agirre, E. (2000). One sense per collocation and genre/topic variations. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora 2000* (pp. 207-215). Retrieved from http://acl.ldc.upenn.edu/W/W00/W00-1326.pdf

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* In C. N. Candlin & R. Carter (Eds.). New York, NY: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. New York, NY: Cambridge University Press.

Meisuo, Zh. (2000). Cohesive features in the expository writing of undergraduates in two Chinese universities. *RELC, 31* (1), 61-95.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, & S. Petch-Tyson (Ed.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp.199-143). Philadelphia, PA: John Benjamins.

Millar, N., & Lehtinen, B. 2008.DIY local learner corpora: Bridging gaps between theory and practice. *The JALT CALL Journal*, *4*(2), pp. 61–72.

Milton, J., & Tsang, E. S. C. (1993). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton & E. S. C. Tsang (Eds.), *Studies in Lexis* (pp. 215-246). Hong Kong: The Hong Kong University of Science and Technology Language Centre.

Mosenthal, J. H. (1983). Cohesion: Problems with talking about text: A brief commentary.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. C. A. Chapelle & S. Hunston (Ed.). New York, NY: Cambridge University Press.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. New York, NY: Cambridge University Press.

Pascual, M., & Unger, L. (2010). Appraisal in research genres: An analysis of grant proposals by Argentinean researchers. *Revista Signos, 43*(73), 261-280.

Pishghadam, R., & Attaran, A. (2013). Rhetorical patterns of argumentation in EFL journals of Persian and English. *International Journal of Research Studies in Language Learning, 2* (1), 81-90. doi: 10.5861/ijrsll.2012.132

Pritchard, R. J. (1980). *A study of the cohesion devices in the good and poor compositions of eleventh graders.* (Unpublished doctoral dissertation). University of Missouri, Columbia.

Radwan, A. (2012). Rhetorical transfer in L2 writing: The role of second language in L2 writing. *Arab World English Journal, 3*(2), 365-399.

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.

Römer, U. & Matthew B. O. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 6*(2), 159-177.

Römer, U. & Wulff, S. (2010). Applying corpus methods to writing research: Explorations of MICUSP. Journal of Writing Research, 2(2), 99-127

Römer, U. (2009). English in academia: Does nativeness matter? *International Journal of English Studies, 20*(2), 89-100.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47.

Sinclair, J. M., & Mauranen, A. (2006). *Linear Unit Grammar*. Amsterdam: John Benjamins.

Trethewey, N. (2010). *Beyond Katrina.* Athens, GA: University of Georgia Press.

Velve, J. R. (2005). Unifying prosody and syntax for right- and left-edge coordinate ellipsis. *Lingua, 115*, 483-502.

White, P. R. R. (2003). Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text, 23(*2), 259-284.

Widdowson, H. G. (1978). *Teaching language as communication.* New York, NY: Oxford University Press.

Williams, J. M. (2010). *Style, lessons in clarity and grace* (10th ed.). G. G. Colomb (Ed.). Boston, MA: Longman.

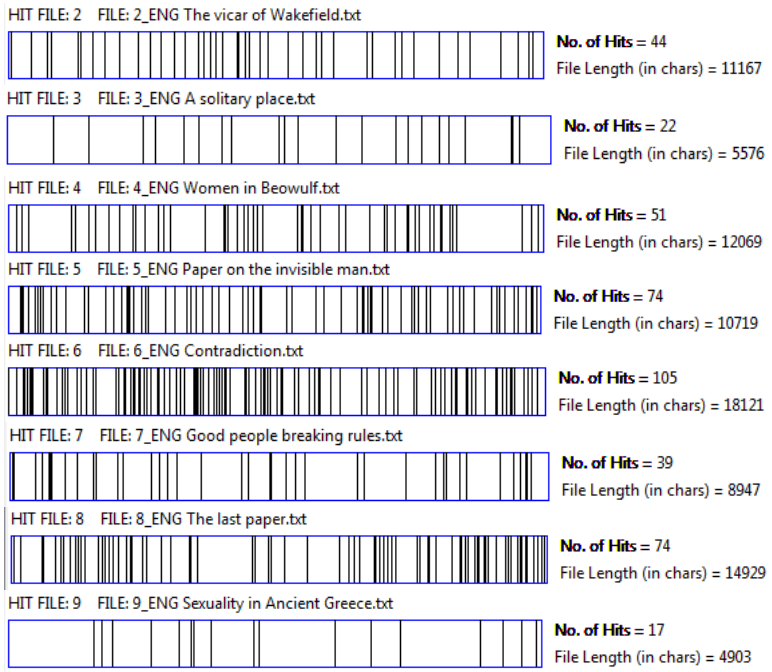
Xi, Y. (2010). Cohesion studies in the past 30 years. *The International Journal – Language Society and Culture,* 31, 139-147.

Yeung, L. (2009). Use and misuse of 'beside': A corpus study comparing native speakers' and learners' English. *System, 37,* 330-342.

Yorkey, R. (1974). Practical EFL techniques for teaching Arabic-speaking students. *International Review of Applied Linguistics in Language Teaching, 37*(4), 291-320.

Appendix

Concordance Plot for Total And-Coordination in Each Essay Across The Three Corpora in terms of Discipline Groups (MICUSP-NNS and MICUSP-NS) and Topics (NUCLE-NNS)

MICUSP-NNS	MICUSP-NS
Arts & Humanities	
	

HIT FILE: 9 FILE: 9_HIS.G1.04.1.txt

No. of Hits = 128
File Length (in chars) = 26111

HIT FILE: 14 FILE: 14_PHI.G0.15.1.txt

No. of Hits = 26
File Length (in chars) = 7000

HIT FILE: 15 FILE: 15_PHI.G0.16.1.txt

No. of Hits = 45
File Length (in chars) = 13159

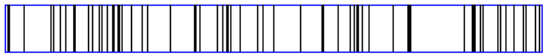
HIT FILE: 16 FILE: 16_PHI.G1.02.1.txt

No. of Hits = 94
File Length (in chars) = 37725

HIT FILE: 17 FILE: 17_PHI.G1.04.1.txt

No. of Hits = 44
File Length (in chars) = 15595

HIT FILE: 10 FILE: 10_ENG The purgatory of the postmodern.txt



No. of Hits = 63
File Length (in chars) = 9480

HIT FILE: 11 FILE: 11_ENG Carwin and the imp of the perverse.txt

No. of Hits = 50
File Length (in chars) = 9140

HIT FILE: 12 FILE: 12_ENG Sport literacy and rhetoric as power.txt

No. of Hits = 294
File Length (in chars) = 37333

HIT FILE: 13 FILE: 13_ENG Domesticity in Cold War black fiction on the left.txt

No. of Hits = 322
File Length (in chars) = 56062

HIT FILE: 14 FILE: 14_CIS The Parthenon Frieze.txt

No. of Hits = 27
File Length (in chars) = 9961

HIT FILE: 15 FILE: 15_CIS Corpse demons in Ancient Greek magic.txt

No. of Hits = 132
File Length (in chars) = 37419

HIT FILE: 16 FILE: 16_HIS Sex education in East and West Germany.txt

No. of Hits = 405
File Length (in chars) = 63495

HIT FILE: 17 FILE: 17_LIN National identity and language education policy.txt

No. of Hits = 129
File Length (in chars) = 22731

HIT FILE: 20 FILE: 20_PHI Emotivism.txt

No. of Hits = 7
File Length (in chars) = 3199

HIT FILE: 21 FILE: 21_PHI Featus rights.txt

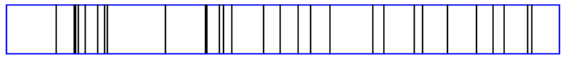
No. of Hits = 38
File Length (in chars) = 14456

HIT FILE: 22 FILE: 22_PHI Pretended ideas in Hume's treatise.txt

No. of Hits = 135
File Length (in chars) = 47329

Life Science

HIT FILE: 1 FILE: 32_BIO.G0.15.1.txt



No. of Hits = 29
File Length (in chars) = 8364

HIT FILE: 10 FILE: 10_NRE.G1.31.1.txt

No. of Hits = 105
File Length (in chars) = 24875

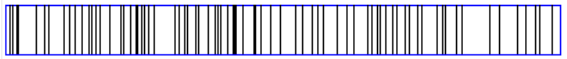
HIT FILE: 11 FILE: 11_NUR.G0.03.1.txt

No. of Hits = 58
File Length (in chars) = 14317

HIT FILE: 12 FILE: 12_NUR.G0.04.1.txt

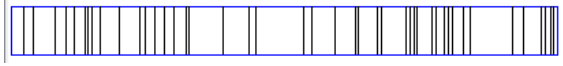
No. of Hits = 80
File Length (in chars) = 16432

HIT FILE: 13 FILE: 13_NUR.G0.07.1.txt



No. of Hits = 80
File Length (in chars) = 17624

HIT FILE: 18 FILE: 18_NRE Natural resources and the environment.txt



No. of Hits = 45
File Length (in chars) = 9436

HIT FILE: 19 FILE: 19_NUR Circumscision.txt

No. of Hits = 114
File Length (in chars) = 30289

Social Science

HIT FILE: 18 FILE: 18_POL.G3.02.1.txt

No. of Hits = 105
File Length (in chars) = 21431

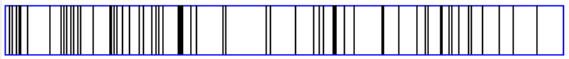
HIT FILE: 19 FILE: 19_PSY.G0.15.1.txt

No. of Hits = 42
File Length (in chars) = 11313

HIT FILE: 20 FILE: 20_PSY.G0.34.2.txt

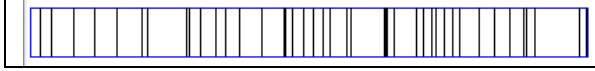
No. of Hits = 41
File Length (in chars) = 10331

HIT FILE: 21 FILE: 21_PSY.G1.11.4.txt



No. of Hits = 63
File Length (in chars) = 12532

HIT FILE: 22 FILE: 22_PSY.G2.10.1.txt



No. of Hits = 42
File Length (in chars) = 11931

HIT FILE: 1 FILE: 32_ECO Illicit drug.txt

No. of Hits = 82
File Length (in chars) = 25370

HIT FILE: 23 FILE: 23_POL Democratization of the EU.txt

No. of Hits = 71
File Length (in chars) = 17933

HIT FILE: 24 FILE: 24_POL The role of party identification in voting behavior.txt

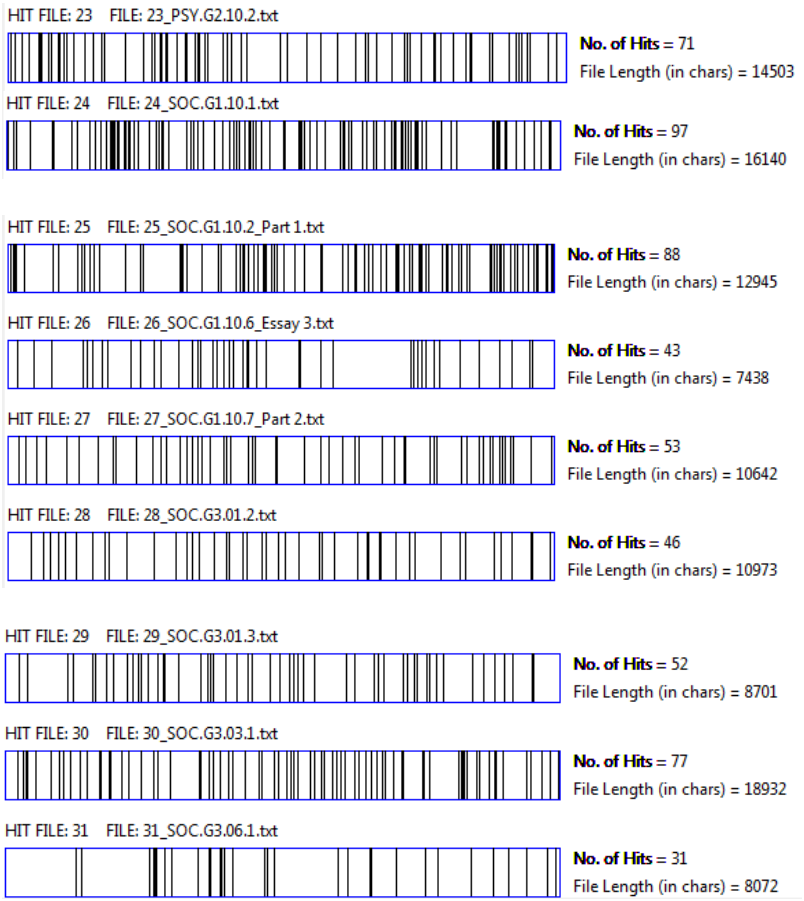
No. of Hits = 57
File Length (in chars) = 19193

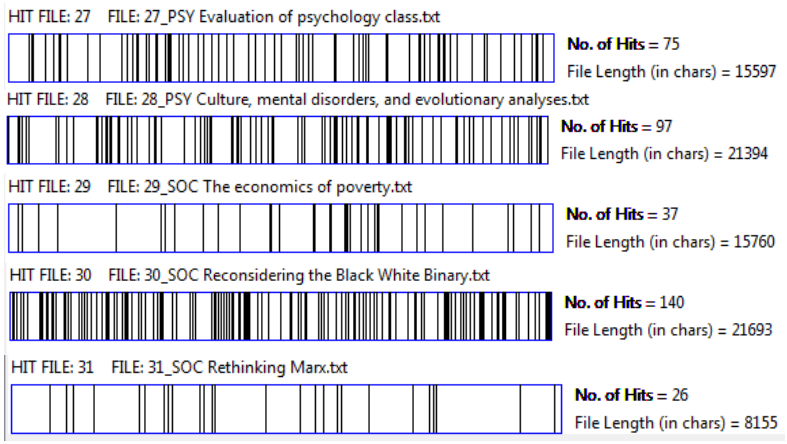
HIT FILE: 25 FILE: 25_POL Evegenia Ginzburge's into the whirlwind.txt

No. of Hits = 67
File Length (in chars) = 19526

HIT FILE: 26 FILE: 26_ POL Measuring racial prejudice.txt

No. of Hits = 73
File Length (in chars) = 13834





NUCLE-NNS	
Topic: *Technological design of 3D printers*	Topic: *Disclosure of genetics testing results*

HIT FILE: 1 FILE: B1_A3.txt

No. of Hits = 5
File Length (in chars) = 3164

Topic: *Surveillance technology*

HIT FILE: 2 FILE: A48.txt

No. of Hits = 10
File Length (in chars) = 2242

HIT FILE: 18 FILE: A768.txt

No. of Hits = 16
File Length (in chars) = 4526

HIT FILE: 29 FILE: A1263.txt

No. of Hits = 11
File Length (in chars) = 2594

HIT FILE: 31 FILE: A1353.txt

No. of Hits = 13
File Length (in chars) = 4036

Topic: *Engineering design*

HIT FILE: 14 FILE: A588.txt

No. of Hits = 3
File Length (in chars) = 2158

HIT FILE: 19 FILE: A813.txt

No. of Hits = 9
File Length (in chars) = 4839

Topic: *Public spending on older adults*

HIT FILE: 4 FILE: A138.txt

No. of Hits = 28
File Length (in chars) = 4143

HIT FILE: 6 FILE: A228.txt

No. of Hits = 22
File Length (in chars) = 4367

HIT FILE: 9 FILE: A363.txt

No. of Hits = 5
File Length (in chars) = 2353

HIT FILE: 10 FILE: A408.txt

No. of Hits = 13
File Length (in chars) = 3646

HIT FILE: 15 FILE: A633.txt

No. of Hits = 20
File Length (in chars) = 4343

HIT FILE: 17 FILE: A723.txt

No. of Hits = 11
File Length (in chars) = 4410

HIT FILE: 16 FILE: A678.txt

No. of Hits = 9
File Length (in chars) = 4930

HIT FILE: 7 FILE: A273.txt

No. of Hits = 20
File Length (in chars) = 5270

HIT FILE: 24 FILE: A1038.txt

No. of Hits = 23
File Length (in chars) = 4247

Topic: *Nuclear power*

HIT FILE: 11 FILE: A453.txt

No. of Hits = 13
File Length (in chars) = 2882

HIT FILE: 13 FILE: A543.txt

No. of Hits = 4
File Length (in chars) = 2867

HIT FILE: 20 FILE: A858.txt

No. of Hits = 26
File Length (in chars) = 4215

HIT FILE: 23 FILE: A993.txt

No. of Hits = 26
File Length (in chars) = 7275

HIT FILE: 12 FILE: A498.txt

No. of Hits = 26
File Length (in chars) = 4754

HIT FILE: 21 FILE: A903.txt

No. of Hits = 14
File Length (in chars) = 3841

HIT FILE: 22 FILE: A948.txt

No. of Hits = 14
File Length (in chars) = 4264

HIT FILE: 25 FILE: A1083.txt

No. of Hits = 13
File Length (in chars) = 4573

HIT FILE: 30 FILE: A1308.txt

No. of Hits = 16
File Length (in chars) = 4466

HIT FILE: 28 FILE: A1218.txt

No. of Hits = 19
File Length (in chars) = 6789

Unclear arguments

HIT FILE: 3 FILE: A93.txt

No. of Hits = 15
File Length (in chars) = 3369

HIT FILE: 5 FILE: A183.txt

No. of Hits = 11
File Length (in chars) = 4757

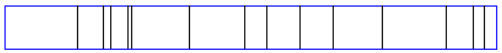
HIT FILE: 8 FILE: A318.txt

No. of Hits = 13
File Length (in chars) = 3553

HIT FILE: 26 FILE: A1128.txt

No. of Hits = 20
File Length (in chars) = 5115

HIT FILE: 27 FILE: A1173.txt



No. of Hits = 14
File Length (in chars) = 4244