SYNTACTIC COMPLEXITY OF EFL, ESL AND ENL: EVIDENCE FROM THE INTERNATIONAL CORPUS NETWORK OF ASIAN LEARNERS OF ENGLISH

DONG QI

(M.A.), GDUFS

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ARTS

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously

_____

ACKNOWLEDGEMENTS

My gratitude goes to a number of people who have helped me in the completion of this thesis.

First of all, I would like to thank my supervisor, Associate Professor Vincent B Y Ooi from National University of Singapore who provided constant guidance, advice and support throughout my entire program. Without his help, this study would never have been completed.

I am also grateful for the two anonymous reviewers who have offered detailed and thought-provoking suggestions for revision. Besides, in the early stage of drafting, Professor Bao Zhiming and Dr. Justina Ong from National University of Singapore offered invaluable comments. Professor Lourdes Ortega from Georgetown University and Professor Yukio Tono from Tokyo University of Foreign Studies also provided help one way or another for my work.

During the data collection, Dr. Yosuke Sato, Dr Chonghyuck Kim and my classmates Lim Ching Geck and Nattadaporn Lertcheva also helped me enrol research participants, for which I am always thankful.

Last but not least, I need to express my heartfelt gratitude to my family members and friends in my home country for their unfailing encouragement and spiritual support.

TABLE OF CONTENTS

SUMMARY

In response to calls for more corpus-based studies at the syntactic level, this study is an attempt to further extend the scope of learner corpus research by investigating the syntactic complexity of EFL, ESL and ENL exemplified by the International Corpus Network of Asian Learners of English (ICNALE). Specifically, based on certain syntactic complexity measures, this study intends to reveal how the language proficiency of the three groups is related to the syntactic complexity measures as shown in their writing, how those measures correlate to each other and how topics influence the syntactic complexity. Three sub-corpora of the ICNALE are employed as the research data, representing the three varietal types respectively. The ICNALE features the strict control over variables such as time, topic and proficiency level, ensuring the maximum reliability of comparison. Data used in this study is both automatically and manually annotated with a detailed multidimensional annotation scheme of syntactic complexity features, aiming to reveal the syntactic information which is unsearchable from raw corpora.

Research findings suggest that global complexity measures and subordination-based complexity measures seem to be stable indicators of proficiency levels. Syntactic complexity features within a certain group are relatively stable, regardless of their proficiency levels. Coordination-based, phrasal and specific complexity measures divided by sentences rather than clauses are generally better indicators of proficiency. T-unit-based measures are disputable in signalling proficiency levels. Correlations between certain measures are also established and explained tentatively. As for the effect of

topic, there seems to be a higher level of syntactic complexity for topic “part-time job” in terms of most measures, supporting the argument that certain topics can induce more complex sentences.

The significance of this study lies in its contribution to revealing the certain features of syntactic complexity of the three groups, which are seldom systematically studied in previous literature due to the lack of strictly controlled corpora. Moreover, based on a relatively detailed annotation scheme, this study also takes the influence of multiple issues like proficiency levels and topic into consideration and offers a clearer picture of how those issues interact with the syntactic complexity across or within the three groups. The research findings might shed light on the following aspects: methodologically, this study illustrates how to use annotated learner corpora to examine the syntactic complexity tentatively; pedagogically, teaching methods and material might be improved accordingly to help learners to approximate native writers in terms of syntactic complexity.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A2_0:	Waystage
B1_1:	Threshold: Lower
B1_2:	Threshold: Upper
B2_0	Vantage or higher
B/C:	Be-copula with Adjective Structures per Clause
B/S:	Be-copula with Adjective Structures per Sentence
CEFR:	The Common European Framework for Reference
CIA:	Contrastive Interlanguage Analysis
CN/C:	Complex Nominals per Clause
CN/S:	Complex Nominals per Sentence
CN/T:	Complex Nominals per T-unit
CP/C:	Coordinate Phrases per Clause
CP/S:	Coordinate Phrases per Sentence
CP/T	Coordinate Phrases per T-unit
C/S:	Clauses per Sentence
C/T:	Clauses per T-unit
CT/T:	Complex T-unit per T-unit
DC/C:	Dependent Clauses per Clause
DC/S:	Dependent Clauses per Sentence
DC/T:	Dependent Clauses per T-unit
EFL:	English as a Foreign Language
ENL:	English as a Native Language
ESL:	English as a Second Language
I/C:	It-cleft Structures per Clause

ICE: The International Corpus of English

ICLE: The International Corpus of Learner English

ICNALE: The International Corpus Network of Asian Learners of English

IRB: The Institutional Review Board

I/S: It-cleft Structures per Sentence

MLC: Mean Length of Clauses

MLS: Mean Length of Sentences

MLT: Mean Length of T-units

POS: Part of Speech

T/S: T-unit per Sentence

VP/T: Verb Phrases per T-unit

VST: Vocabulary Size Test

CHAPTER ONE: INTRODUCTION

1.1 Introduction

Syntactic complexity, which is also referred to as "syntactic maturity" or "linguistic complexity", is identified as greater variety of sentence patterns, or progressively more elaborate language (Foster & Skehan, 1996, p. 303). Given its importance and difficulty, syntactic complexity has been extensively studied in the field of second language acquisition (SLA) and first language acquisition in the past decades. In corpus linguistics, it was not until the early 1990s that some corpus linguists tentatively studied learners' syntactic patterns with a heavy reliance on SLA theories and practices. Notably, in corpus linguistics, much has been published on lexical issues of language, covering a wide range of research topics in various backgrounds. As pointed out by some linguists (e.g. Granger, 2009; Tono, 2010), however, there is a relative lack of attention on the syntactic information of language production in corpus linguistics, partially due to the difficulty of extracting such information from corpora (Gilquin, 2003). Such a scarcity is especially true when it comes to corpus-based comparison of EFL learners, ESL learners and ENL learners: most existing studies only focus on the language production by a certain language group or two groups. Moreover, among those corpus-based studies on language production at sentence level, it is not difficult to spot some limitations in certain aspects such as the selection of corpora and measures for analysis. Further corpus-based studies on syntactic complexity of the three groups based on comparable datasets are necessary in this regard.

Based on three highly comparable sub-corpora from the ICNALE (Ishikawa, 2011), this study intends to explore how syntactic complexity is related to the proficiency of EFL, ESL and ENL, how certain syntactic complexity measures correlate with others and how topic influences syntactic complexity. During the construction of various components of the ICNALE, writing conditions such as time constraints, topics and availability of references were strictly controlled, making those sub-corpora as homogenous and comparable as possible. Besides, for those EFL and ENL components, different proficiency levels are assigned with a unified framework called the Common European Framework of Reference (CEFR) (Little, 2007), providing a strong support for establishing the link of proficiency and certain syntactic complexity measures. Meanwhile, for the native writer component, both novice native writers and expert native writers are evenly distributed and identified, taking the influence of writing expertise on syntactic complexity into consideration. All corpus data used in this study is annotated with a detailed multidimensional scheme of syntactic complexity features, making in-depth analysis and comparisons possible.

1.2 Thesis organization

Consistent with the research objectives, this thesis is organized as follows: Chapter one outlines the research topic and motivation for the study before offering the background of this research and syntactic complexity measures used in this study, pointing out how the existing studies can be improved or extended and affirming the necessity of this research. Based on the implications drawn from chapter one, the second chapter deals with the research design, in which the rationale of the design, research questions and

data construction/annotation are detailed. In the third chapter, the data analysis is presented to demonstrate the findings of this research and answer each research question, followed by a discussion of those findings in chapter four. The last chapter concludes the thesis and points out the research directions for further research.

1.3 Research motivation

1.3.1 Importance of syntactic complexity

Being able to employ various sentence patterns is an indispensable writing skill for successful writers. This issue is often translated into the syntactic complexity of writing. Syntactic complexity has been long observed by many linguists and language teachers, who have paid special attention to the contribution of those more complex sentence patterns in expressing complex ideas and improving writing quality. It is acknowledged that "certain syntactic structures, such as subordinate clauses, relative clauses, and complex noun phrases allow writers to express more complex ideas" (Beers & Nagy, 2011, p. 184). In this respect, using complex sentence patterns is necessary for clearly stating one's ideas effectively. In addition, the use of complex grammatical structures signals effective writing (de Haan & van Esch, 2006; Reilly, Zamora, & McGivern, 2005; Rimmer, 2008; Schleppegrell, 2004). Complex sentence structures are thus related to the quality of writing in this connection.

On the contrary, simple sentences are often regarded to show the weakness of learners. Many linguists and educators regard them an important disadvantage in writing and argue that they may result in the deduction of writing scores (e.g. Davidson, 1991; Hamp-Lyons, 1991; Reid,

1993; Vaughan, 1991). Among many others, Hinkel (2003) conducted a qualitative analysis of writing by over 1000 learners and native speakers, noticing that those learners employed excessively simple syntactic constructions. Such a heavy reliance on simple sentence patterns and difficulty of using more complex sentence patterns may be attributable to the current mainstream teaching method in writing instructions. According to Connors (2000), recent writing instructions tend to focus on some higher level stages of writing process such as planning and revising, and consequently the 'syntax of writing' is given less attention. Clearly, variation of different sentence patterns, especially the employment of more complex sentence patterns, is critical for good writings when it comes to English learners, who may have difficulty in using various English sentence patterns at ease.

1.3.2 Scarcity of corpus-based studies on sentences

Despite the importance and difficulty of using more complex sentences for learners, studies at sentence level in corpus linguistics are less common compared with those studies on lexical issues, not to mention studies on the syntactic complexity. It seems that syntactic complexity is generally examined in SLA research instead. In SLA research where learner corpora have gradually gained popularity, syntactic complexity is more often than not explored without the use of corpora. Most of those SLA studies are based on experiments, tapping the production of learners' writing (e.g. Foster & Skehan, 1996). Those experiments generally provide three major types of data: "Language use data, metalingual judgments and self-report data" (Ellis, 1994, p. 670). The difficulty of drawing firm conclusions from a narrow

empirical basis is underlined by many SLA and corpus linguists. Among others, Gass and Selinker (2008, p. 55) argue that it is “difficult to know with any degree of certainty whether the results obtained are applicable only to the one or two learners studied, or whether they are indeed characteristic of a wide range of subjects”. Learner corpus research features “a wider empirical basis than has ever previously been available” (Granger & Paquot, 2009, p. 16) is thus adopted to study the syntactic complexity in this research.

Acknowledging the advantage of learner corpus research over traditional SLA research in providing a wider range of empirical basis, linguists also need to note that the potential of learner corpora to study the syntactic complexity of learners has not yet been fully realized. The scarcity of corpus-based studies on sentence patterns is largely because of the difficulty of extracting such information with appropriate corpora/tools (Gilquin, 2003). Moreover, “the background of corpus research largely rooted in the European tradition of descriptive and functional linguistics” (Tono, 2010, p. 9) also contributes to this scarcity. On one hand, querying of raw corpora is still limited to the search of lexical information. Obviously, words are easier to count and classify than sentence structures (Rimmer, 2008). Although certain parsed corpora can be used to study certain characteristics of sentence patterns, they are not always available to the public. On the other hand, while various computational tools for analysing corpus have been devised globally in the past decades, most of them are seldom used to examine the syntactic features, except for a few of them such as Hawkins and Buttery (2010), Lu (2010) and Saville (2010).

The scarcity of corpus-based studies on sentences is especially true in the comparison of EFL, ESL and ENL in a single study. Among them, studies on the use of sentences by ESL learners such as Singapore English learners are also not very common. Undeniably, language acquisition in Singapore with a context of complex multilingual settings deserves special attention (Kirkpatrick, 2011). As noted by Schneider (2007: 157), the syntax of Singapore English features many distinctive rules and patterns; however, they are seldom systematically examined based on learner data. Among those existing studies where syntactic features of Singapore English are discussed, we may still find relatively small datasets by researchers with a tendency to emphasize colloquial Singapore English (e.g. Deterding, 2010; Low & Brown, 2005) rather than the type of 'standard' Singapore English described by Low (2010), not to mention the written English used by Singapore English learners. Given the scarcity of corpus-based studies on sentences, especially the comparison of EFL, ESL and ENL in a single study, the current research aims to bridge this gap by conducting a corpus-based project to examine the syntactic complexity of writings by EFL learner, ESL learners (Singapore English learners here) and ENL writers.

1.4 Literature review

1.4.1 Overview of studies on syntactic complexity in L2 study

Syntactic complexity, as the major approach to study sentence variation, has been explored in a wide range of areas in applied linguistics including first language acquisition, language disorder studies and SLA research. As for its applications in SLA research, existing studies can be grouped into the following categories: First, syntactic complexity often

refers to evaluating the impact of different experiment settings on language production, for instance, the impact of planning time on language production (Foster & Skehan, 1996). Besides, syntactic complexity is also applied to study the variation of language production across language groups, for example, the language production of eight learner groups with different first language (Taguchi, Crawford, & Wetzel, 2013). Third, syntactic complexity has also been applied to map the proficiency levels within certain learner groups, for instance, the study of the relationship between Chinese English learners' language proficiency and syntactic complexity measures (Lu, 2011).

Generally, syntactic complexity has been explored through the calculation of the average length of certain syntactic units, density of subordination and frequency of certain linguistically more complex forms (Ortega, 2012). Wolfe-Quintero, Inagaki, and Kim (1998) and Ortega (2003) offer two research syntheses of studies on syntactic complexity, in which various existing studies are compared and evaluated. Notably, subsequent studies on syntactic complexity have seldom been systematically reviewed and compared. In what follows, some representative newer studies on syntactic complexity are thus reviewed with an emphasis on four critical issues related to the study: 1). measures for studying L2 syntactic complexity; 2) reliability of those measures; 3). the relationship between L2 proficiency level and syntactic complexity; and 4) the automatic analysis of L2 syntactic complexity and manual annotation.

1.4.2 Measures for studying syntactic complexity

A number of representative measures for syntactic complexity are summarized in Table 1. Consistent with the scope of this research, only those measures used for L2 writing studies are included. Despite the advances of knowledge on syntactic complexity, those measures for examining syntactic complexity do not really change much compared with those used in the past decades, except for the integration of some specific forms as measures for syntactic complexity. Regarding the selection of those measures, two points merit discussion here: the first is on the persistence of T-unit-based measures in those studies and the second is on the integration of new measures.

Among those measures illustrated in Table 1, measures with T-unit calculated have gained popularity among existing studies since several decades ago. Such popularity is especially true for the mean length of T-units, which is used as the most widespread measure for syntactic complexity (e.g. Armstrong, 2010; Brown, Iwashita, & McNamara, 2005; Larsen-Freeman, 2006; Nelson & Van Meter, 2007). T-unit, the minimal terminable unit, was first proposed by Hunt (1965), who defined it as "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" (Hunt, 1970, p. 4). Hunt (ibid) argued that mean length of T-units and clauses per T-unit, together with words per clause were the three most reliable indicators of syntactic complexity. After that, this argument has been supported by the overwhelming majority of researchers in the follow decades. In the two early research syntheses on syntactic complexity by Wolfe-Quintero et al. (1998) and Ortega (2003), they agree on that this measure serves as the most reliable measure for discriminating proficiency

levels based on their review of over 40 studies in total. Even in some new studies, mean length of T-unit is still used as the major measure for discriminating syntactic complexity.

Although T-unit is widely applied in various studies on sentence complexity in the past decades, its plausibility is questioned by some linguists (e.g. Bardovi-Harlig, 1992; Biber, Gray, & Poonpon, 2011; Foster & Skehan, 1996; Gaies, 1980; Lu, 2011). Their criticism can be grouped into the following categories. First, by "imposing uniformity of length and complexity on output that is not present in the original language sample" (Bardovi-Harlig, 1992, p. 391), T-unit may distort the original intentions of language learners who produce sentences rather than T-units. Second, a T-unit analysis ignores some useful information such as the coordination (Ortega, 2012) and noun clausal features embedded in noun phrases (Biber et al., 2011), both of which are also important indicators of syntactic complexity for certain group of learners. Third, some empirical studies have found that T-unit measures are not always capable of differentiating syntactic complexity because those more proficient learners are not necessarily those who produce longer T-units in (e.g. Smart & Crawford, 2009). It is also noted that there is not any theoretical rationale for the use of T-unit.

Apart from the first two categories of measures, the third category of measures which features the specific forms of language production seems to be neglected by most researchers in their studies of syntactic complexity. Knowing the length of production of unit and subordination does not necessitate a full understanding of syntactic complexity because the first two categories of measures can only provide certain quantitative information

which is not so helpful for making specific inferences or judgments. In certain cases, following measures from the first two categories without careful consideration may result in the misinterpretation of data. Length does not necessarily increase as those learners progress to more advanced levels. It is possible for more advanced learners to produce longer T-units, however, such an increase can be a result of increased use of complex phrases such as coordinate phrases and complex nominals, rather than increased use of subordination (Lu, 2010). Likewise, advanced learners may also choose to use more embedding rather than longer syntactic structures, resulting in shorter production units (Arthur, 1979; Kern & Schultz, 1992). In this regard, other more specific measures are needed to complement the length-based measures and subordination-based measures.

Complementing or extending the first two mainstream categories of measures, other types of measures targeting at certain characteristics of syntactic complexity are of great importance given the possible limitations of the first two categories. The integration of some other types of forms to measures for syntactic complexity may help researchers further reveal certain characteristics of syntactic complexity (e.g. Lu, 2011; Vyatkina, 2013). Notably, the integration of those forms has its empirical support in some L2 studies. For instance, features such as phrasal features and complex nominals can further contribute to the in-depth exploration of syntactic complexity. Phrasal features are found to index writing quality and are thus recommended to be incorporated into the measure for syntactic complexity (Biber & Gray, 2010; Biber et al., 2011; McNamara, Crossley, & McCarthy, 2010; Rimmer, 2006). Complex nominals often serve as an alternative to

relative clauses (Hundt, Denison, & Schneider, 2012) and may also reflect the complexity of sentences (Gordon, Hendrick, & Johnson, 2004; Halliday, 1989; Halliday & Webster, 2004). In a comparison of syntactic complexity features of academic writing and spoken language, Biber et al. (2011) find that "complex nominals (rather than clause constituents) and complex phrases (rather than clauses) are common in academic writing", both of which are generally considered to be less grammatically complex. Such an observation refutes the assumption that more subordination structures equal more grammatically complex sentences, which makes those syntactic complexity studies purely based on subordination-related measures self-contradictory.

Those measures featuring certain forms of syntactic complexity are certainly not limited to those mentioned in Table 1. Extension or further justification of them in future research is still necessary since those measures related to phrasal complexity and complex nominals are still relative new in the research into syntactic complexity. Compared with length-based measures and subordination-based measures, those measures are relatively less frequent in previous studies. They are more specific compared with the complexity measures based on lengths of certain units or subordination structures. As observed by some linguists, the more specific a measure is, the more revealing it is (Hudson, 2009). Notably, while length-based measures and subordination-based measures have long enjoyed popularity in syntactic complexity research, those specific complexity measures also begin to gain popularity in some latest studies, which may help us gain a clearer picture of how syntactic complexity is represented and evaluated.

Table 1 Selected measures for examining syntactic complexity in the past ten years (2004-2013)

Category of measures	Measures	Sources
Length-based measures	Mean length of sentences	Benedikt Szmrecsanyi (2004)
	Mean length of T-units	Armstrong (2010)
	Mean length of clauses	Byrnes (2009)
Subordination-based measures	Mean number of clauses per T-unit	Becker (2010)
	Mean number of dependent clauses per clauses	Wigglesworth and Storch (2009)
	Frequency of dependent clauses	Biber et al. (2011)
	Frequency of subordinate conjunction	Vyatkina (2012)
Specific forms of syntactic complexity	Frequency of tenses, modal verbs and voices (passive forms)	Ellis and Yuan (2005)
	Frequency of coordinate structures, complex nominal structures and non-finite verb structures	Vyatkina (2013)
	Frequency of phrasal features such as Post–noun-modifying prepositional phrase	Taguchi et al. (2013)

1.4.3 Reliability of the studies on syntactic complexity

The reliability of corpus-based studies is often undermined due to the inappropriate selection of measures and sometimes due to the undesirable statistical methods. When using those measures for studying syntactic complexity in their studies, researchers seldom justify the reliability of those measures. Acknowledging the possible application of syntactic complexity measures for studying language, researchers also need to attach importance to the reliability issues of those measures and think twice before selecting measures of syntactic complexity. Notably, some measures are too abstract and general to reveal the language phenomenon and thus failing to reveal some information specifically. Such a limitation is especially true when only one or two measures are used to study the syntactic complexity of sentences, including some quite new studies, for instance, Vaezi and Kafshgar (2012) applied only two measures, average sentence length and ratio of subordination to study syntactic complexity of writing. Syntactic complexity is a complicated multi-faceted phenomenon, and it is thus problematic to use only one or two measures to examine such a construct in language production (Biber et al., 2011; Myhill, 2006; Rimmer, 2008). Pointing out the limitation of relying on only one or two measures does not mean that researchers need to employ as many measures as possible. Some studies employing various measures are actually using redundant measure because some of their measures are examining exactly the same thing (Beers & Nagy, 2009; Norris & Ortega, 2009). From what has been covered on the reliability of those measures, we need to draw a lesson that a wide range of measures is necessary to ensure the reliability of syntactic complexity analysis while

redundant measures should be removed to make the analysis is more productive.

Another critical issue regarding the reliability of those corpus-based studies is on the statistical methods for analysing data. Some researchers tend to treat each learner group as a whole without considering the individual difference among each group, which is one of the central themes in SLA research (e.g. Dornyei, 2005). Durrant and Schmitt (2009, p. 168) note that comparing corpora as wholes may neglect the individual differences of learners and may therefore potentially produce misleading results. Certainly, comparison of averages is not always meaningful in the analysis "because averages often obscure the distribution of frequencies in the sample" (Hinkel, 2003). Flowerdew (2010) also notices the discrepancies between the frequencies based on the whole data and means of frequencies based on individual texts, realizing that there may be greater idiosyncratic variations in the learners' use which should be emphasized in future research. Appropriate statistical methods are thus necessary to bridge the methodological gap, for instance, t-test can be used to describe the individual differences of individuals. Those individual differences should be studied qualitatively to complement the corpus findings if necessary. As noted by Reinhardt (2010, p. 95), "a mixed corpus and qualitative approach to the analysis of learner language" should be employed to ensure the individual features are also considered.

1.4.4 Syntactic complexity and proficiency

It is very common for researchers to equate syntactic complexity with proficiency level directly. The link between certain syntactic measures

and proficiency is taken for granted in some studies. For instance, subordination in writing is considered to be more complex than coordination (e.g. Bardovi-Harlig, 1992; Carter & McCarthy, 2006; Hopper & Traugott, 2003; Purpura, 2004; Willis, 2003). However, as suggested in some studies (e.g. Bardovi-Harlig & Bofman, 1989; Beers & Nagy, 2009, 2011; Gaies, 1980; Osborne, 2011; Song, 2006; Taguchi et al., 2013), the correlation between certain syntactic complexity measures and writing proficiency is not necessarily strong. Notably, the development of discourse and sociolinguistic repertoires is also necessary for the development of proficiency (Ortega, 2003). Certainly, complex sentences do not always equal good sentences because measures for syntactic complexity do not always translate into valid measures of writing proficiency or quality (Lu, 2011). In some situations, complex sentences:

> "can be awkward, convoluted, even unintelligible…Conversely, relatively simple sentences can make their point succinctly and emphatically. Often, of course, sentence variety is best" (Weaver, 1996, p. 130).

It is of paramount importance to note that different measures can "serve different interpretive purposes for different proficiency levels" (Norris & Ortega, 2009, p. 573). For instance, intermediate learners may use more subordination structures when they begin to progress to advanced learners. However, when they have become advanced learners, they may also use more complex nominals to replace those subordination structures in order to meet the requirement of academic English. To summarize, "the ability to produce complex sentences is probably best understood as a

necessary but not sufficient condition for writing high quality texts" (Beers & Nagy, 2009, p. 187).

1.4.5 Automation of syntactic analysis vs. manual annotation

Automatic analysis of syntactic information is appealing to corpus linguists; however, such systems are still far from being perfect due to the difficulty of extracting syntactic structures efficiently and exhaustively (Gilquin, 2003). Employment of measures calculated automatically may invite the issue of software accuracy (e.g., Vyatkina, 2012), and such an issue is especially serious when it comes to learner data that often contains various kinds of errors. If we have known that the accuracy rate of parsing tools is not as high as Part of Speech (POS) taggers, we may consider employing a POS tagger. However, those POS taggers are almost all based on the annotation scheme developed for native speakers, consequently, the reliability of their application on learner data lacks empirical evidence (Dıaz-Negrillo et al2010; Dickinson & Ragheb, 2009), for instance, the correlation between human rater and automatic method of syntactic complexity is quite low, only 0.49 correlation value in Miao and Klaus's case (2011). This dilemma can explain why automatic systems for analysing the syntactic complexity of first language are more common than those used for analysing second language.

Nevertheless, some latest automatic tools seem to be quite useful in analysing syntactic complexity by learners. Lu (2010, 2011) devised a pioneering automatic system to examine the syntactic complexity of learners' written language based on the Stanford Parser (Klein & Manning, 2003) and Tregex (Levy & Andrew, 2006). According to Lu (ibid), this automatic tool

is quite reliable because of result and the manual annotation matches quite well. In Lu's study (2011), number of complex nominals per clause, mean length of clauses and mean length of sentences were found to be the best discriminators for different proficiency levels. Undeniably, such automatic systems do have their advantages of processing a large quantity of texts at the same time and incorporating comprehensive measures. To further complement it, manual intervention or even manual annotation for certain measures is still necessary for obtaining reliable and exhaustive information retrieval when automatic annotation does not guarantee the full analysis.

1.5 Syntactic complexity used in this study: A multidimensional annotation scheme of syntactic complexity

Consistent with the scope of this study, a multidimensional annotation scheme is proposed for the data annotation following the recommendation by Norris and Ortega (2009): 1) General complexity, 2) complexity via subordination, 3) complexity via coordination, 4) complexity via phrasal elaboration 5) and other specific measures of syntactic complexity. In addition, due to the disputable role of T-unit-based measures in signalling syntactic complexity (see section 1.4.2), they will be put into the sixth category. Before moving on to the description of those measures, the introduction to units used for annotation is in order.

1.5.1 Introduction of units

Sentence: A sentence is defined as "a string of words with a capital letter at the beginning of the first word and a period or another terminal punctuation mark after the last word" (Homburg, 1984, pp. 91-92). Identifying a sentence is "straightforward in the written language" (Crystal,

2008, p. 432), because punctuation is considered as a helpful indicator of sentencehood.

Clause: “Clause is a term used in some models of grammar to refer to a unit of grammatical organization smaller than the sentence, but larger than phrases, words or morphemes” (Crystal, 2008, p. 78). As for the composition, “a clause is a grammatical unit that includes, at minimum, a predicate and an explicit or implied subject, and expresses a proposition” (Hartmann & Stork, 1972, p. 137). It includes independent clauses, adjective clauses, adverbial clauses, and nominal clauses.

Dependent clause: A dependent clause is often called a subordinate clause. It is defined as “a clause that is embedded as a constituent of a matrix sentence and that functions like a noun, adjective, or adverb in the resultant complex sentence” (Quirk, Greenbaum, Leech, Svartvik, & Crystal, 1985, p. 44).

Coordinate phrase: Coordinate phrases are phrases linked together by conjunctions “that link constituents without syntactically subordinating one to other” (Hartmann & Stork, 1972, p. 54).

Complex nominal: Cooper’s study (1976) categorized complex nominals into two types: complex nominals with heads or without heads, however, this thesis only counts on those noun phrases with heads. Specifically, complex nominals include (1) nouns plus adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; (2) nominal clauses; and (3) gerunds and infinitives in subject, but not object position (ibid).

Be-copula structures with predicative adjectives: In this sentence structure, “be” is used as a copula to link the subject and the predicative adjective. Such a syntactic structure is proved to be a characteristic of simple structures by less proficient learners (Hinkel, 2003), and thus it is incorporated as a measure for syntactic complexity.

It-cleft structure: This sentence structure is composed of a pronoun “it” and a form of the verb be, optionally accompanied by the negator “not” or an adverb, followed by the specially focused element (Biber, 959).

T-unit: T-units. A T-unit is “one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it” (Hunt 1970: 4).

1.5.2 Global complexity

Global complexity measure, or general complexity, aims to give a basic quantitative description of sentence. In this study, sentence rather than T-unit is selected as the basic unit of language production because of the limitations of T-units revealed by many studies (e.g. Bardovi-Harlig, 1992; Biber et al., 2011; Foster & Skehan, 1996; Gaies, 1980). Sentence is easier to calculate and it is arguably regarded to reflect the direct choices of learners. Moreover, total clauses per sentence may further reveal the general information of sentences and it is thus also regarded as the second global syntactic complexity measure in this research.

1.5.3 Complexity by subordination

In this research, measure of subordination is based on the calculation of dependent clauses. More specifically, ratios between dependent clauses and total clauses/total sentences are calculated to mirror the subordination in

sentences. It is assumed that subordination may signal more advanced writing compared with coordination.

1.5.4 Complexity by coordination

Coordination is generally regarded to be indicative of less complex syntactic structures because the relations between the structures are much easier to master for less proficient learners compared with subordination. In this regard, coordination seems to be more frequent in less proficient learners who may have difficulty in using more subordination structures in their writing. In this research, coordination phrases are identified and calculated against the total number of clauses and total number of sentences in each text.

1.5.5 Phrasal complexity

A few linguists have realized the contribution of phrasal complexity to syntactic complexity (e.g. Biber et al., 2011) although phrasal features are not extensively studied in most studies on syntactic complexity. In this research, the length of clauses is examined first because the complexification of phrases will always increase the length of clause indirectly. It is noted that phrasal complexity measures are not studied exhaustively in this research due to the concern of feasibility and the scope of this research. Instead, only complex noun phrases (complex nominals) are studied here. Other categories of phrases like verb phrases and preposition phrases are thus excluded in the annotation and further analysis.

1.5.6 Specific measures of syntactic complexity.

While the previous four categories all focus on certain features of syntactic complexity that can be automatically identified, the fifth category

of measures may call for manual identification. The rationale to use the two pair of measures is largely based on the observation by Hinkel (2003) who found that frequent use of be-copula with adjective structures was considered to be a feature of less advanced learners while the use of it-cleft structures was often a characteristic of advanced writers. The first two measures in this category deal with the characteristics of "simple" syntactic patterns, more specifically, "be-copula" with adjective structures. I hypothesize that they will be overused by those less proficient learner groups in the study, say, EFL learners. Adopting the other two measures is a straightforward decision: "it-cleft" structure is generally considered to be more difficult and it is expected to discriminate learners across proficiency levels and native speakers.

1.5.7 T-unit-based complexity

Due to the disputable role of T-unit-based measures in signalling syntactic complexity (see section 1.4.2), they will be studied in a category alone in the scheme. The eight T-unit-related measures are Mean length of T-units (MLT), Verb Phrases per T-unit (VP/T), Clauses per T-unit (C/T), Dependent Clauses per T-unit (DC/T), T-unit per Sentence (T/S), Complex T-unit per T-unit (CT/T), Coordinate Phrases per T-unit (CP/T) and Complex Nominals per T-unit (CN/T).

Table 2 presents the syntactic complexity measures and the way of calculation for the thesis. This detailed multidimensional annotation scheme aims to provide a clear picture of syntactic complexity in EFL learners, EFL learners and ENL writers, allowing more fine-grained comparisons and qualitative analysis. Although corpus linguistics is mostly quantitative in

nature, qualitative analysis based on a detailed scheme of those features is still necessary because it is pointless to say “use thing less often” without knowing what the relevant alternatives would be in specific contexts (Hunston, 2002, p. 209). In the follow analysis, qualitative information will be provided when necessary to complement the quantitative findings. Offering rich information about the language use at sentential level, a detailed multidimensional annotation scheme can shed invaluable light on the research.

Table 2 Syntactic complexity measures used in the study

Category	Measures	Calculation	Code
Global complexity	Mean length of sentences	Words/Sentences	MLS
	Clauses per sentence	Clauses/Sentences	C/S
Complexity by subordination	Dependent clauses per clause	DC/Clauses	DC/C
	Dependent clauses per sentence	DC/Sentences	DC/S
Complexity by coordination	Coordinate phrases per clause	CP/Clauses	CP/C
	Coordinate phrases per sentence	CC/Sentences	CP/S
Phrasal complexity	Mean length of clause	Words/Clauses	MLC
	Complex nominals per clause	CN/Clauses	CN/C
	Complex nominals per sentence	CN/Sentences	CN/S
Specific complexity features	Be-copula structures per clause	B/Clauses	B/C
	Be-copula structures per sentence	B/Sentences	B/S

	It-cleft structures per clause	I/Clauses	I/C
	It-cleft structures per sentence	I/Sentences	I/S
T-unit-based complexity features	Mean length of T-units	Words/T-unit	MLT
	Verb Phrases per T-unit	VP/T-unit	VP/T
	Clauses per T-unit	Clauses/T-unit	C/T
	Dependent Clauses per T-unit	DC/T-unit	DC/T
	T-unit per Sentence	T-unit/Sentences	T/S
	Complex T-unit per T-unit	C T-unit/T-unit	CT/T
	Coordinate Phrases per T-unit	CP/T-unit	CP/T
	Complex Nominals per T-unit	CN/T-unit	CN/T

1.6 Chapter conclusion

In consideration of the importance of syntactic complexity for quality writing, more corpus-based studies on syntactic complexity is necessary. Despite the advances of studies on syntactic complexity, there is still plenty room for further improvement with regard to their research design. The selection of appropriate measures and the reliability of research design merit special attention in future research. Besides, linking proficiency level to syntactic complexity blindly may distort the research result. Finally, while automatic annotation is very efficient in processing certain aspects of language, manual annotation is still necessary for studying certain syntactic features of learners' language in future research. In this study, both automatic and manual annotation methods are employed. The former is used to compute a large number of indices which has already proved to be quite reliable in Lu's study (2010) while the latter targets selected certain features of syntactic complexity to ensure feasibility and accuracy of manual work provided the analysis is statistically meaningful and reliable.

CHAPTER TWO: RESEARCH DESIGN

2.1 Introduction

Given the importance of syntactic complexity and the scarcity of corpus-based studies on syntactic issues, this corpus-based study positions itself to bridge this gap by investigating the syntactic complexity of EFL learner, EFL learners and ENL writers jointly. Three sub-corpora of the ICNALE are employed as the research data, including the Singapore Component (a typical ESL learner group in multilingual settings), ENL component and China component (a typical EFL learner group). Composed of timed writing by learners and native speakers with the same two topics, the ICNALE features the strict control over corpus construction to maximize comparability. Unlike most previous cross-sectional corpus-based studies where proficiency levels of certain groups are not seriously considered, this study has applied the CEFR to map the proficiency levels of participants in each group in an attempt to conduct more reliable comparison within learner groups. Additionally, ENL component of this corpus is further divided into the novice native writer part and expert native writer part, making more refined comparisons of expert and trainee native writers possible. With a detailed multidimensional scheme of syntactic complexity features mentioned in chapter one, all samples of the research data are annotated to afford more detailed analysis.

Before moving on to the introduction to the other issues of research design, the explanation of the rationale for this research design is in order. After that, the research scope is delimited, followed by the introduction of research questions and account of the data composition.

2.2 Rationale of the research design

First of all, this study is a corpus-based study on syntactic complexity which was generally explored in SLA research. Notably, studies on sentential issues are much less compared with those on lexical issues in corpus linguistics while most SLA researchers are inclined to base their studies of syntactic complexity on experiments. Such a discrepancy may raise a question that why corpora rather than experiments should be used to study syntactic complexity in this the research. This question can be resolved through the introduction of Contrastive Interlanguage Analysis (CIA) (Granger, 1994), which shows the distinctive advantage of learner corpus research in this issue.

Besides, unlike most existing corpus-based studies on sentence patterns where only the target learner data is included, this study has incorporated a native writer sub-corpus and both EFL learner sub-corpus and ESL learner sub-corpus for reference. Thanks to the strict control over various variables such as time, topic and length when constructing the corpora, the three datasets used in this study allow high level of comparability, which is not always attainable in other studies where many variables are beyond control. The purpose of comparing learner data and native data is straightforward because native data can provide benchmark for learners and tell researchers how different learners are from native speakers. Besides, comparing different learner data, e.g. ESL data and EFL data, may contribute to a better understanding of the language progression in interlanguage system.

2.2.1 Contrastive Interlanguage Analysis in learner corpus research

In this research, CIA based on learner corpora is chosen as the research method. Unlike traditional contrastive analysis where different languages are compared, CIA concerns varieties of the same language. It "involves quantitative and qualitative comparisons between native language and learner language (L1 vs. L2) and between different varieties of interlanguage (L2 vs. L2)" (Granger, Dagneaux, Meunier, & Paquot, 2009, p. 18). Figure 1 illustrates the bidirectional comparisons of CIA.

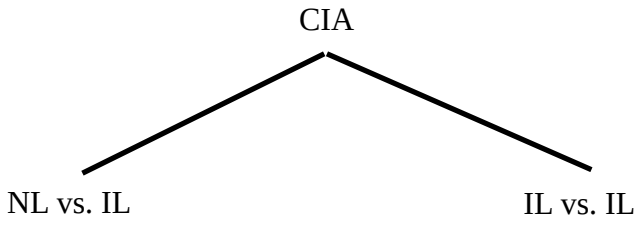


Figure 1 Contrastive Interlanguage Model

Since the early 1990s, learner corpora have gained popularity among both corpus linguists and SLA linguists. Despite the wide application of learner corpora in SLA, "learner corpus research has not yet fully realized its potential as its links with SLA have been somewhat weak" (Granger, 2009). This is especially true when it comes to the study of sentences whereas learner corpora are assumed to be an excellent basis for studying grammatical complexity (ibid). Learner corpora, "one of the most important resources for studying interlanguage (Borin & Prutz, 2004), can record sizeable authentic language use by L2 learners, shedding invaluable light on how L2 learners acquire and use language (Granger, 2009; Tono, 2009a). Moreover, learner corpora can test "the findings previously made on the basis of limited data of a small number of informants and generalize their

findings" (Xiao, 2007). Last, the information extracted from learner corpora can help construct computational model of SLA theories with attested language use data (Tono, 2009b).

While certain advantages of learner corpora over traditional SLA experiment are acknowledged, researchers also need to note some distinctive merits of SLA research, for instance, the complexity measures and the theories in SLA research can be applied to the learner corpus research given the "inherently interdisciplinary nature of learner corpus research" (Granger, 2009, p. 14).

2.2.2 Comparison of syntactic complexity of EFL and ESL learners

It is also noted that despite the wide coverage of both varietal types respectively, systematic comparisons between EFL, ESL and ENL are not common (Davydova, 2012; Nesselhauf, 2009; Van Rooy, 2011), much less on the syntactic aspects. A systematic comparison of the three groups of data can contribute to a better understanding of how language users from the three groups differ from one another. However, due to the lack of available reference corpora where variables are strictly controlled to ensure comparability, most existing corpus-based studies on L2 writing only deal with a certain group of language users, i.e., target learner group (e.g. Taguchi et al., 2013). In some other cases, the reference corpora used in their studies seem to be lack of reliability because the composition of those reference corpora is quite different from that of the original ones. Some researchers have realized it and may try to compromise it. For instance, Laporte (2012) compared the use of "make" in the International Corpus of Learner English (The ICLE) and a small part of the International Corpus of English (ICE)

(student writing and exam scripts) to examine the differences of “make” in EFL and ESL varietal types. The problem is, due to the composition of ICE, the portion suitable for making comparisons with the ICLE is quite small, only around 40,000 words in each sub-corpus. This may consequently influence the representativeness in comparison. The current study benefits from the strict control over various variables such as time, topic and length during the construction of corpora. With the three highly comparable sub-corpora including representative varietal types of ENL, ESL and EFL, high level of comparability is realizable in the data comparison.

2.3 Scope of measurement

Target measures of syntactic complexity used in this study fall into six categories. The first five were recommended by Norris and Ortega (2009): 1) General complexity, 2) complexity via subordination, 3) complexity via coordination, 4) complexity via phrasal elaboration 5) and specific measures of syntactic complexity. The sixth category consists of the disputable T-unit-based complexity measures. Measures from the six categories are supposed to constitute a multidimensional coverage of syntactic complexity features. While the first two categories dealing with length-based units and density of dependency are common in previous studies on syntactic complexity, the following three categories may provide some fine-grained information of syntactic complexity. Coordination might be used more often by less advanced learners generally whereas phrasal elaboration seems to be a feature of advanced writing and more formal writing like academic writing. In this regard, they seem to be indicative of proficiency of writing. The last second category of measures is devoted to

those specific forms which may reflect the variation of forms in accordance with the acquisitional timing. Variation in accordance with the acquisitional timing seems to be more of the nature of L1 acquisition, however. Given the emphasis on L2 writing and the nature of the data (argumentative writing), measures in the fifth category should be selected with caution. Apart from being suitable for the analysis of L2 writing, they should be able to index features of syntactic complexity and preferably have been tested in previous studies. After careful consideration, occurrences of be-copula and it-cleft as recommended by Hinkel (2003) have been manually annotated in this research to serve as specific features of syntactic complexity. The last category is for disputable T-unit-based measures.

2.4 Research questions

After the discussion on the rationale and scope of this research, three research questions are presented to address the key issues of this research topic, covering 1) the relationship between proficiency level and syntactic complexity for participants from ESL (Singapore), EFL (China) and ENL backgrounds, 2) How do different complexity measure correlate with each other for the three groups, 3) the influence of topic on syntactic complexity for the three groups.

2.4.1 Relationship between proficiency level and syntactic complexity

The first research question intends to establish the possible links between the proficiency levels of those participants and syntactic complexity measures. While previous studies have varying opinions on the correlation between proficiency level and syntactic complexity, the current study intends to answer this question with a relatively larger size of comparable data.

Research question 1: What is the relationship between syntactic complexity and proficiency level of the three groups as a whole/ respectively?

It is assumed that due to the nature and proficiency of the three groups, their relationship between proficiency level and syntactic complexity may not follow a linear line. In other words, those syntactic complexity measures signalling proficiency levels may be different for the three groups. For instance, for learners of lower proficiency, coordination based-measures may be a better indicator of them while for those expert native writers the frequent use of complex nominals may be one of their characteristics. A more qualitative analysis is conducted to further identify the complexity features of data by manually identifying be-copula and it-cleft structures, representing both features of simplistic writing and more advanced writing as suggested by Hinkel (2003).

It is noted that the sixth category of complexity measures, T-unit-based measures will only be covered in discussions related to this research question due to the scope and depth of research.

2.4.2 Correlation between different syntactic complexity measures

Since sentence is the basic unit of writing and the variation of other syntactic complexity measures may always influence it, it is reasonable to assume that certain syntactic complexity measures may correlate with it or with other measures.

Research Question 2: How do different measures of syntactic complexity correlate with each other to realize complexification among the three groups of participants?

By understanding the correlation of those measures, we can get a better understanding of how the three groups differ from each other by establishing the possible connections between those measures. Accordingly, some pedagogical suggestions can be made based on the result analysis.

2.4.3 Influence of topic on syntactic complexity

Benefiting from the strict control over variables in the corpus construction, the two topics used to elicit writing from participants can help us reveal the influence of topic on syntactic complexity. In some earlier studies, topics in corpora were found to account for the differences between varietal types (Danzak, 2011; M. Hundt & Vogel, 2011; Wulff & Römer, 2009). As revealed in the findings from Danzak (2011), significant differences in syntactic information of writing were generally based on the topic on the writing sample. Given the two distinctive topics used during the corpus construction, it is possible to take the influence of topic into consideration when analysing the syntactic complexity of the three groups.

Question 3: Is there any effect of topic on syntactic complexity for ESL learners' writing as compared to those of the EFL learners and ENL writers? If so, in what way does topic influence syntactic complexity features?

The influence of topic on the syntactic complexity might be an interesting and promising research direction. If certain topics are found to be able to induce more syntactically complex sentence patterns, teachers can use them more to help learners improve their syntactic complexity in a more effective manner.

2.5 Data construction

In order to address the research questions raised above, the selection of the most appropriate data is of paramount importance. The decision to select the ICNALE as the data for the study merits explanation first. After that, a brief introduction to the ICNALE is presented to illustrate its suitability for this study, followed by a description of the compilation process for the Singapore component.

2.5.1 Decision on data selection

The quality of corpora where the evidence about language acquisition is based on is a prerequisite for learner corpus research (Tomasello & Stahl, 2004) since the quality of the corpus will largely decide whether the corpus findings are reliable and whether there will be some new observations. Before making decision on choosing an existing corpus or making a new corpus for the study, I considered the following factors and tried to strike a balance between them: 1) size and representativeness issues and 2) control over variables and availability of reference corpora.

2.5.1.1 Size and representativeness of corpora

For general corpus, especially those corpora of native language, the size is of great importance. Nevertheless, for learner corpora, size is not necessarily a decisive factor for its value. Granger (2009, p. 17) observes that:

"Big is not necessarily beautiful…the SLA specialist attaches more importance to control over the many variables that affect learner production than to sheer size. As a result, learner corpora need to be assembled on the

basis of very strict design criteria and a wide range of variables should ideally be recorded for each learner production."

The pursuit of size for corpus research is primarily because of the assumption that large corpora can be more representative and small corpora are generally less representative of language. The problem is, due to the availability of learner data, the vast majority of learner corpus studies are based on relatively small corpora. The concern over size for learner corpora should give way to the concern over representativeness, which plays a more important role compared with sheer size. While the size of a learner corpus is generally not as large as native corpora, the number of contributors to the corpus data would be more critical for deciding the representativeness. Assume there are two corpora of the same size, say, one million. If the first one million is composed of 1000 learners' works while the second is composed of 2000 learners' works, the latter should be more representative since there are more participants. The "direct relation between the size counted in number of words and representativeness measured in number of learners" (Granger, 2011, p. 9) does not hold true for learner corpus.

Obviously, the small-scale corpus has the following advantages: (1) high comparability in terms of variables, and (2) possibility of fully manual analysis (Laporte, 2012). Moreover, if the number of participants of the data is large enough, the representativeness of learner corpora can still be guaranteed.

2.5.1.2 Control over variables and availability of reference corpora

Due to the limited availability of learner corpora, many existing studies are unable to exert strict control over variables. This is especially true

when researchers want to compare their learner group with a native group. Researchers have to compromise in order to find a relatively acceptable reference corpus in most cases.

Moreover, proficiency and writing expertise should also be given due attention when choosing a reference corpus (Hasselgård & Johansson, 2012), or the results derived from the analysis may actually be because of the proficiency difference rather than of other causes. As emphasized by Hasselgård and Johansson (ibid), the research objective and learners' situation should determine whether professional native speaker corpora or learner native speaker corpora should be used. Thus it is important to bear in mind that the distinction between expert native writers and learner native writers should be made in making comparisons. On one hand, control over variables such as time, genre and length in learner corpus research is critical for approaching comparability. On the other hand, in order to make more fine-grained comparison, both the novice and expert native writer should be included in the research data if they are available, because adopting expert native writers only may "set too high a standard" for examining learners' writings (Hyland & Milton, 1997; Lorenz, 1999; McCrostie, 2008). For the study which focuses on the differences of syntactic complexity between EFL, ESL and ENL, comparable reference corpora should be sought in order to identify the differences and answer the research questions.

As proposed by Myles (2005), "researchers need to make sure that the corpora they use are adapted to the research agendas, rather than adapting research questions to the corpora readily available". In order to provide data for the thesis, I undertook the construction of the Singapore

component of the ICNALE under the guidance of my supervisor. In the remaining part of this section, some basic information of the ICNALE and the construction of the Singapore ICNALE are introduced.

2.5.2 Introduction to the ICNALE

Given those factors influencing the decision on corpora for the study, the ICNALE seems to be a desirable option for the study because it well strikes a balance between those factors. The ICNALE is a collection of 1.3 million words of essays written by 2,600 college students in 10 Asian countries and areas plus 200 English native speakers (Ishikawa, 2013, p. 94). The size of the ICNALE is supposed to be large enough for studying learner language, especially for the syntactic features in this study, which generally do not require a very large dataset compared with those studies on lexical issues. Likewise, the number of participants for the ICNALE may also suffice the need for realizing representativeness. Moreover, since the ICNALE also exerts strict control over many other variables such as time and topic, it is especially appropriate for the study which involves detailed comparison with controllable variables.

It is well-known the size of corpora is an important concern for evaluating the validity of them, because if the size is too small, it is "difficult to know with any degree of certainty whether the results obtained are applicable only to the one or two learners studied, or whether they are indeed characteristic of a wide range of subjects" (Granger, 2011, p. 31). Although the corpus size of the ICNALE is not as large as some of the other learner corpora like the ICLE (Granger et al., 2009), the number of participants

involved is still large enough. On the whole, the representativeness of the ICNALE is quite satisfying.

Variables including genre, topic, time limit, availability of references and proficiency are strictly controlled during the compilation of the ICNALE, providing a solid basis for detailed comparison. Unlike some other learner corpora where there may be a mixture of genres, all the samples of the ICNALE are argumentative writing. Such control over genre intends to minimize the uncontrollable variables in order to make more reliable comparisons possible because genre or register may decide the grammar of writing (Beers & Nagy, 2009, 2011; Biber, 1999). A recent experiment indicates that "the relationships between syntactic complexity and text quality are dependent both on the genre of the text and the measure of syntactic complexity used" (Beers & Nagy, 2009). This supports the need for controlling the genre of writing in order to make the corpus composition homogeneous.

In order to approach the maximum comparability, the essay topics are also controlled. In this study, there are two topics in this research:

(A) "It is important for college students to have a part time job."

(B) "Smoking should be completely banned at all the restaurants in the country."

Each participant was required to write two short articles around 200 to 300 words for each of the two topics. Given the significant effect of topic on the language production (Danzak, 2011), the "rationale for choosing the essay title" (Rimmer, 2008, p. 31) should be validated here. Both topics are expected to elicit highly personalized response from participants because

“the language sample can be a valid indicator of accomplishment in the grammatical structures of interest” (Purpura, 2004, p. 233).

Another important feature of the ICNALE is that proficiency level of each learner participant is labelled with the external criteria based on CEFR. Given the heterogeneity of the second language learner population, chronological age or other issues like grade level should not be considered as reliable discriminators of learner proficiency (Gaies, 1980). Such a classification of proficiency level based on external criteria features is definitely more reliable than the categorization of learners in some studies where internal criteria features like age and grade level were applied. Moreover, identifying proficiency levels of participants in larger corpora would provide more insight into their differences and facilitate analysis (M. Hundt & Vogel, 2011). Only when the proficiency levels of participants are taken into consideration can the conclusion of differences between different varietal types be meaningful (Carlsen, 2012; M. Hundt & Vogel, 2011; Tono, 2009b; Wulff & Römer, 2009). For native data, the distinction between trainee native writers who are students and expert native writers who are professionals is also drawn in the ICNALE, thus incorporating expertise of writing as a controllable segment in proficiency cline.

Compared with the ICLE, which is the most popular corpus among learner corpus research, the ICNALE has its advantages in strict control over variables.

In the ICLE corpora, timed and untimed essays are not strictly balanced in number and many studies tend to treat them as one category only (Hundt &Vogel, 2011). Besides, the availability of references in the ICLE is

not controlled. In the ICNALE, on the contrary, each participant is given 20 to 40 minutes for the writing without using references like dictionary or the Internet.

Table 3 provides a comparison of the ICNALE and the ICLE in order to illustrate the differences of them and the advantages of the ICLE for the current study. From this table, it is possible to find that the ICNALE excels in the comparability because of its strict control over those variables. It is noted that such a corpus with strict control over variables is rare in corpus research.

On the whole, the ICNALE has a satisfying size for learner corpus research with enough participants to ensure representativeness. The genre and even topic used in the ICNALE are also strictly controlled to ensure comparability. Moreover, time allowed for participants and availability of references are also determined at the compilation stage, further controlling the variables that might influence the result of analysis. Last, the proficiency levels of learners and distinction between native students and native professionals are also identified, making refined comparisons possible.

Table 3 Comparison of the ICNALE and the ICLE

	The ICNALE	The ICLE
Size (total)	1.3 million	3.7 million
Size (Sub-corpora)	~90,000-200,000 words	~200,000-500,000 words
Average length of writing	200-300 (±10%) words	~700 words
Participants per sub-corpus	100-400	~330
Control over genre	+ Argumentative	-(Argumentative & literary essays)
Control over topic	+ (two topics)	-
Control over time	+ (20~40 minutes)	(65% were uncontrolled)
Availability of references	-	(65% were uncontrolled)
Identification of proficiency	+ (CEFR)	-

Three sub-corpora of the ICNALE are employed in this study after careful consideration since they can represent the typical language user groups of EFL, ESL and ENL. The three sub-corpora are the Singapore Component (a typical ESL learner group in multilingual settings), ENL component and China component (a typical EFL learner group) of the ICNALE. The basic information of the three sub-corpora can be found in Table 4. A detailed account of the construction of Singapore component will be offered in the next section.

Comparison of EFL data and ESL data with ENL being their benchmark is necessary because there are some shared features of EFL and ESL (e.g. Gilquin & Granger, 2011) as well as some distinctive features in each varietal type (e.g. B. Szmrecsanyi & Kortmann, 2011) awaiting further exploration. Moreover, comparison can also be made within each varietal type given the proficiency levels involved in each group. The fine-grained comparison may help reveal how the syntactic knowledge of learners

progress in the interlanguage system, which can be used to propose a theoretical model to mirror the progression process and be applied to the improvement of teaching material or teaching methods. The composition of the native sub-corpus as a reference corpus deserves a mention here for its even distribution of novice native writer part (trainee) and expert native part (expert).

Table 4 Composition of corpora in the study

Variety	Participants/Essays	Proficiency	Tokens
ESL (Singapore)	200/ 400	B1_2; B2_2	96,733
EFL (China)	400/ 800	A1_2; A2_1; B1_2; B2_0	194,613
ENL	200/ 400	Trainee/Expert	88,792

2.5.3 Construction of the Singapore ICNALE

The construction of Singapore component of the ICNALE took around three months (supervised by A/P Professor Vincent Ooi and executed by the author). After obtaining the approval from Institutional Review Board (IRB), posters were put up online to enrol eligible Singapore participants. Participants were limited to those undergraduates born and raised in Singapore. In response to the requirements of the IRB, ethical considerations were given before enrolling participants. All participants joined this project willingly without coercion. They were told the basic requirements for participating in the project and those who did not meet the enrolment requirements were rejected at the very beginning. All participants agreed to contribute their writing and questionnaire for research purpose. The privacy of participants was strictly protected during the whole process. By the end of

corpus compilation, over 220 participants contributed to the data, 200 of which were chosen as the final data for Singapore component of the ICNALE. Apart from the control over other variables like topic and length, writing conditions were also controlled, lest the uncontrolled writing would "confuse the difference in writing conditions with that of writer groups" (Ädel, 2008). Each participant was required to download the Excel file from the website made for this project and complete the tasks in the file on computer. The reason why computer rather than paper was used as the writing media in this research is primarily because computer can facilitate the writing of learners (Li, 2006; Pennington, 2003). According to Pennington (2003), learners may feel more comfortable when they are writing on computer and it is perceived such a writing condition can help researchers elicit more authentic language use. Writing on computer can also facilitate the data processing and save a lot of time because transcription is not necessary for the computerized writing. Last, writing on computer can also reduce the possibility of typos which is beyond the research scope of this study.

In the Excel file downloaded from the website for the Singapore ICNALE, there was also a questionnaire to tap the basic information, language-related information and the vocabulary size of participants. Basic information and language-related information of participants could help the researcher reveal certain characteristics of participants and interpret research findings while the vocabulary test could be used to establish a link between learners' language proficiency and vocabulary size with the CEFR. In other words, the writers' personal characteristics, L2 proficiency, L2 learning

background, and experiences can be investigated in as much detail as possible (Ishikawa, 2013) and thus providing complementary information for analysis. Apart from filling out some language learning background information, participants were required to take an "English vocabulary size test (VST)" (Nation & Beglar, 2007). The project leader Ishikawa (2013, p. 98) argues that VST is "robustly correlated with the general L2" proficiency based on the correlation study of VST score and the English proficiency test score provided in questionnaires of participants. To sum up, the use of questionnaire can contribute to the overall quality of the ICNALE since it can provide additional information of learners which can be used to interpret or even triangulate the research findings.

2.6 Data annotation

Annotation information may greatly facilitate the querying of certain linguistic information (e.g. Dıaz-Negrillo et al., 2010; Meurers, 2005; Meurers & Müller, 2009). In this regard, the annotated corpora are promising because researchers can extend from analyses based on words to a more abstract level of linguistic patterns in language production (Granger, Kraif, Ponton, Antoniadis, & Zampa, 2007; Meurers & Müller, 2009; Vyatkina, 2012). However, most existing learner corpora are raw corpora without much added information. The application of computer tools for POS tagging or parsing English has to some extent liberated the researchers from manual labour of coding such information. Notably, we need to note that almost all of those tools were originally designed for analysing native English. Learners' language production, on the contrary, is not always suitable for the automatic coding with those parsing or tagging tools. Largely because of the

nature of learner language, automatic parsing tools do not always work well on learner corpora. As warned by (Granger, 2009), learner corpus researchers have to be careful with most of these tools based on native speaker data because they are not fully adapted for processing learner data. Previous studies have reported that due to the errors of learner language, the accuracy rate of many quantitative measures may be affected.

In this research, both automatic and manual annotation methods are employed. The automatic method is based on the L2 Syntactic Complexity Analyzer (Lu, 2010) which can automatically count certain measures of syntactic complexity. More specifically, structures like sentences, clauses, coordinate phrases and complex nominals are identified with this system. This can save a lot of time and ensure consistency because the identification of those structures can achieve high computer-annotator agreement, although it is unable to extract the specific measures of syntactic complexity for this thesis. To complement the automatic annotation, a certain amount of manual annotation is conducted tentatively given the relatively small size of the learner corpora. Manual annotation is "time-consuming, but nevertheless the most effective approach available" (Flowerdew, 2010, p. 38). After finishing the annotation, "the annotated information can subsequently be used as search criteria to retrieve all the occurrences in the corpus that match a particular query" (Granger, 2011). Given the necessity of a detailed annotation to further revealing the originally unsearchable information in corpus, the computational tools for both automatic and manual annotation are introduced in the following discussion.

2.6.1 Automatic annotation tool: L2 Syntactic Complexity Analyzer

The automatic system for identifying certain components of sentences can save a lot of time and ensure consistency. According to the designer of L2 Syntactic Complexity Analyser (Lu, 2010), identification of components like sentence length, clause length and number of complex nominals by this system has been checked against the identification by human annotators with a very high level of system-annotator agreement (0.851~1). This suggests that it is quite applicable to count those structures with this software package.

To further test the applicability of this software package, 30 samples, 10 from each sub-corpora, were randomly selected from the research data for manual annotation of structures involved in the current annotation scheme, namely, clauses, complex nominals, dependent clauses and coordinate phrases. The number of structures found in each sample is compared with the number of structures produced in this automatic annotation software package. Table 5 shows the system-annotator agreement of the manual annotation and automatic annotation, supporting the reliability of this tool. According to the statistics, the correlation values of clauses, dependent clauses and coordinate phrases are quite high while the value for complex nominals is relatively low, although on the whole it is still quite satisfying.

Table 5 System-annotator agreement between manual annotation and software annotation on random samples

	Clause	CN	DC	CP
System-annotator agreement	0.973	0.853	0.970	0.975

Given the satisfying identification of those units, this software package is employed to conduct the identification of sentence, clause,

dependent clause, coordinate phrase and complex nominals while the identification of those specific syntactic complexity measures (be-copula with adjective structures and it-cleft structures) will be done through manual annotation, which will be covered in the following section. Based on the occurrences of those structures, values for the syntactic complexity measures for this research are calculated for analysis.

2.6.2 Manual annotation tool: UAM CorpusTool

Given the importance of manual annotation for this learner corpus research on syntactic complexity and the coverage of the multidimensional annotation scheme described above, an appropriate annotation tool should be sought to code the two specific measures of syntactic complexity.

UAM CorpusTool 3.0 (O'Donnell, 2013) was chosen as the manual annotation tool for this study because of its convenience in coding both document information and certain segment information. The manual annotation process is greatly facilitated by dragging the mouse over a certain part of text and matching it with a certain feature stipulated by the researcher. Another advantage of UAM CorpusTool is that it allows semi-auto-coding by assigning new features to one layer of features that have been annotated already or to certain segments that contain a specific string of words. Finally, basic statistics can be performed on this tool, presenting various statistic comparisons of certain annotated features within or between groups as required by the researcher. This can further provide some quantitative information of the data.

With the help of UAM CorpusTool, be-copula with adjective structures and it-cleft structures related to specific measures of syntactic

complexity are annotated in accordance with the multidimensional annotation scheme of syntactic complexity features. Annotator is supposed to follow different layers of the scheme in manual annotation in order to ensure consistency. The semi-automatic annotation is conducted only when the accuracy can be guaranteed. Such a semi-automatic annotation can save considerable time when annotating the native writer data. However, due to the nature of learner language, the automatic annotation of learner data is conducted with special caution, especially for those EFL and ESL learners.

The fact that the researcher and the annotator is the same person may have both its strength and disadvantage. On one hand, the researcher who has designed the annotation scheme is quite familiar with the scheme and is supposed to be efficient of coding data. On the other hand, it is possible that the subjectivity of the researcher may negatively influence the objective annotation process. In order to counter the threat of subjectivity, the annotator is supposed to conduct reliability check on the stratified random samples of the annotated corpus data. In case of disagreement on certain features, the annotator shall check the problem carefully and decide the correct annotation. By doing so, the reliability of manual annotation can be ensured.

The follow two text excerpts illustrate how be-copula with adjective structures and it-cleft structures are annotated manually for this research.

“Recently, there has been a discussion about whether it <u>is important</u> for college students to have a part-time job. There are two opinions about this question. Some people think it <u>is good</u> to have part-time job. But some

other people don't think it is good to do it." (Excerpt of be-copula with adjective structures from CHN_PTJ_024_A2_0.txt)

"For this reason, it is my belief that this dying breed should respect all non-smokers and not subject us to the dangerous consequences of being around cigarette smoke." (Excerpt of it-cleft structure from corpus text ENS_SMK_105_XX_0.txt)

2.7 Chapter conclusion

This chapter begins with the rationale of this research and delimits the research scope. The application of CIA provides a support for making comparisons among the three groups. Among them, the comparisons between EFL, ESL and ENL data are especially meaningful since the findings can help learners to realize how to approximate native writers. After introducing the rationale, three research questions are proposed, focusing on the main topic on this research. The answers to those questions are based on relatively detailed data analysis, which heavily relies on the careful data construction and annotation with the multi-dimensional annotation scheme of syntactic complexity. The ICNALE featuring the strict control over variables is thus selected as the research data for this study, maximising comparability and reliability. Both automatic annotation and manual annotation methods are applied to the research data.

CHAPTER THREE: DATA ANALYSIS

3.1 Introduction

To answer the three research questions, the data processed with L2 Syntactic Complexity Analyzer and UAM CorpusTool is subjected to detailed statistical analysis in accordance with the scheme of syntactic complexity. Those measures are used to examine both the syntactic complexity of the three groups as a whole and within each group respectively. As mentioned earlier, the four proficiency levels of all those EFL and ESL learner participants have been identified with CEFR and the group of native writers is divided into expert part and trainee part. By doing so the proficiency cline ranging from lower intermediate EFL learners to expert native writer has been established, facilitating the detailed comparisons of different complexity measures with other independent variables in line with the research design. In addition to establishing the possible links between proficiency levels and certain syntactic measures, the correlation between certain syntactic complexity measures is also tentatively explored in order to further reveal how syntactic complexity is realized and how the findings can be applied in pedagogy, followed by an examination of the effect of topic on syntactic complexity measures among the three groups as a whole and respectively.

The analysis is based on the observation of those syntactic complexity features of the three sub-corpora of the ICNALE, representing EFL, ESL and ENL group respectively. Following the detailed multidimensional annotation scheme, key features related to syntactic complexity are identified in each text for further statistical analysis, resulting

in the statistics of number of sentences, words, clauses, dependent clauses, coordinate phrases, complex nominals, be-copula structures and it-cleft structures for each sample. Based on the occurrences of them, the 13 syntactic complexity measures of the annotation scheme are computed for each sample, followed by the multi-dimensional comparisons within or across the three groups with other variables.

3.2 Syntactic complexity and proficiency

Proficiency in this research is loosely defined as the writing ability of learners. Syntactic complexity is thus regarded as a reflection of writing ability in syntactical aspect. In other words, a subset of proficiency. Since the proficiency levels of learners in the corpus data have been identified with CEFR and the distinction between student native writers (trainee native writers) and professional native writers (expert native writers) has also been marked, it is reasonable to conceptualize a cline of proficiency. It is believed that in this cline three groups of participants have varying proficiencies. Within each of the two learners' groups, proficiency levels were identified earlier with CEFR. For native participants, a distinction between trainee writers and expert writers was also established during the corpus construction.

Figure 2 illustrates this cline visually. EFL is placed to be the least proficient end of this cline, followed by ESL in the middle of this cline. Naturally, ENL situates at the most proficient end. It is noted that there is an overlapping of proficiency between EFL and ESL since both of them have proficiency levels of B1_2 and B2_0 according to the CEFR identification

during the corpus composition, which may provide added information on comparing EFL learners and ESL learners with the same proficiency levels.

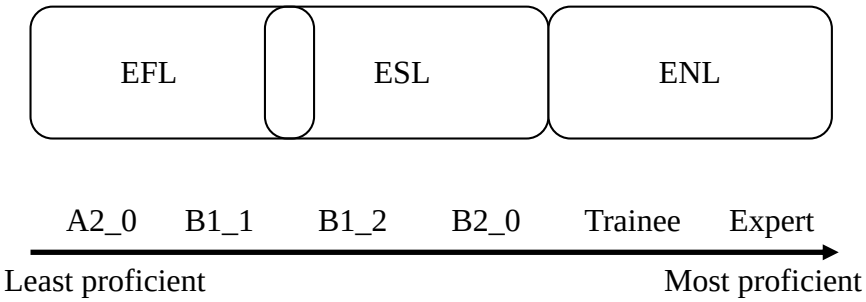


Figure 2 Cline of proficiency in EFL, ESL and ENL

Among linguists (e.g., Lu, 2011: 45), there is an assumption that if certain measures of syntactic complexity, e.g., length-based measures, are found to progress in a way significantly related to the proficiency cline of the three groups, such measures are supposed to be useful indicators of language proficiency in the three groups.

3.2.1 Global complexity measures and proficiency

According to the annotation scheme, global complexity is measured in terms of average sentence length and ratio of clauses per sentence. The first step of analysis is to check if the differences between the three groups are statistically significant. ANOVA tests are performed accordingly. Among each of the three groups, p-values for both measures are smaller than 0.001, supporting the argument that the three groups are statistically different. It is expected that their proficiency levels will follow a cline from EFL to ENL with ESL in the interim of this cline. After that, descriptive statistics is performed on the data to calculate the mean and standard deviation of the three groups, as has been done on the other data analyses in this research. Table 6 suggests that there are significant differences between the three groups in their mean sentence length and number of clauses per sentence,

indicating a strong increase of syntactic complexity from EFL to ENL in terms of the two global syntactic complexity measures. For instance, mean length of sentences for EFL is 16.45 words while the figure for ESL reaches a much larger number of 22.27. For ENL, the figure is even larger, i.e., 25.70, more than 9 words, or one half in total than that of EFL group. Besides, the increasing standard deviation of the three groups further indicates that compared with EFL and ESL learners, ENL writers tend to show more variation in their sentence length and clauses per sentence. This is most probably because learners are always abided by certain rules in writing and focus on forms rather than meanings whereas native writers have a much larger repertoire of techniques to express their ideas freely and do not strictly follow specific rules in their writing.

Table 6 Global complexity measures of EFL, ESL and ENL

Measures	Group	N	Mean	Std. Deviation
MLS	EFL	800	16.45	3.79
	ESL	400	22.27	4.98
	ENL	400	25.70	5.91
C/S	EFL	800	1.89	0.50
	ESL	400	2.19	0.53
	ENL	400	3.06	0.94

Apart from the obvious differences between proficiency levels associated with the language background (EFL, ESL or ENL), the proficiency levels identified with CEFR within EFL and ESL groups, together with the distinction between student native writers and professional native writers, can provide a clearer picture of how proficiency levels are

related to global syntactic complexity measures. A closer examination of the global syntactic complexity measures seems to suggest that what discriminate the three groups of learners are actually not proficiency levels within each group but their linguistic backgrounds across groups: participants from a certain group seem to exhibit similar level of global syntactic complexity, regardless of their proficiency levels. Figure 3 shows that within a certain group, global syntactic complexity measures do not seem to change much while participants' proficiency/writing expertise within each certain group is increasing from the left end to the right end. This is especially true for EFL and ESL learners. Such contradiction might be explained with the linguistic backgrounds of those participants, which can be further explored in future research.

As shown in the Figure 3, for both EFL and ESL groups, their sentence length is largely related to their respective language backgrounds, i.e., EFL or ESL. While there are four proficiency levels in EFL group, the mean length of sentences does not change much from the lowest proficiency level A2_0 to highest learner level B2_0. In the same manner, B1_2 and B2_0 in ESL group do not show much variation.

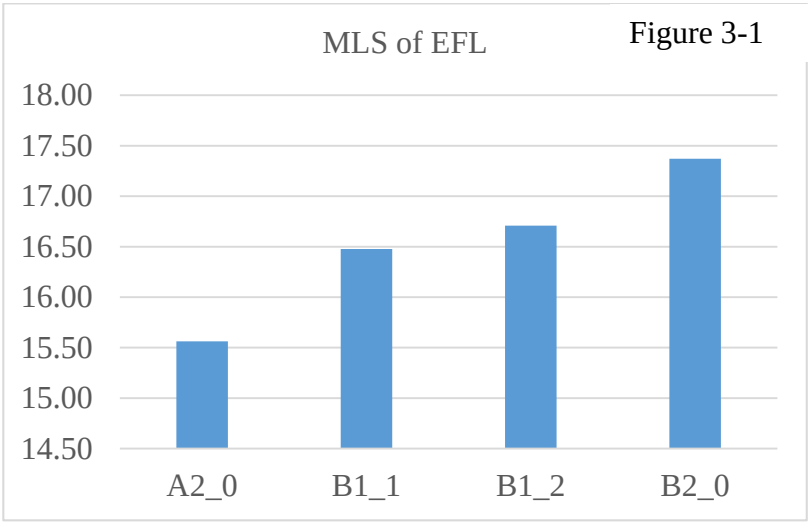


Figure 3-1

Figure 3-2

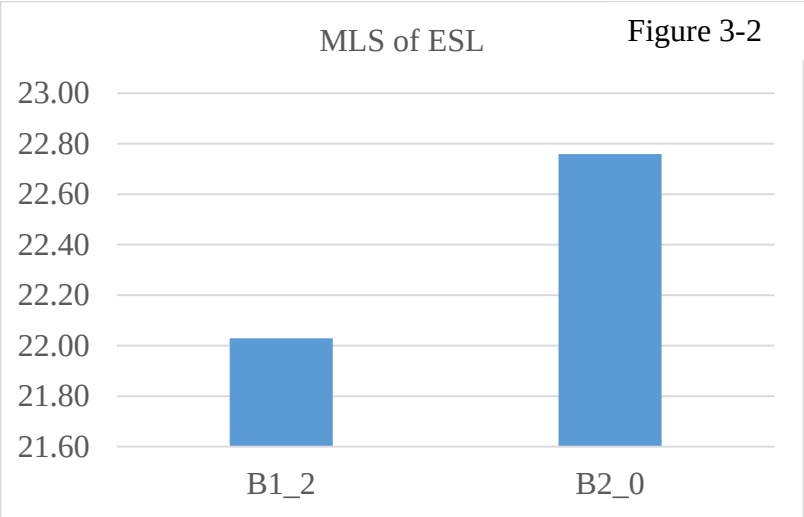
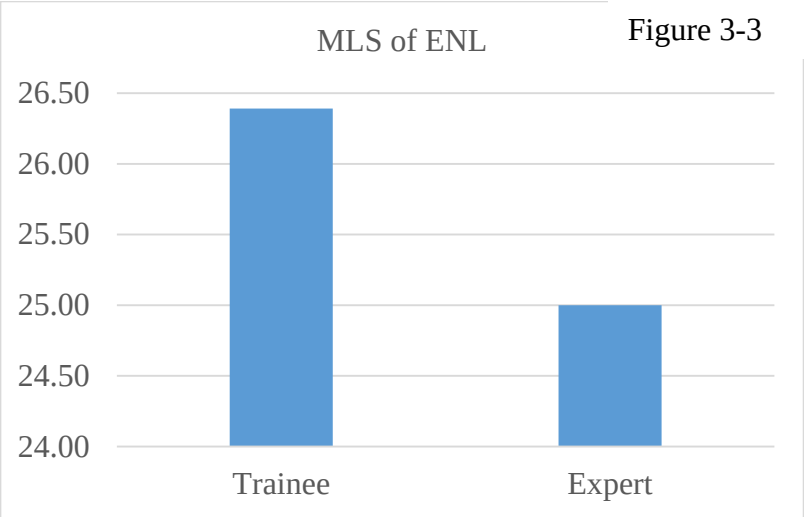


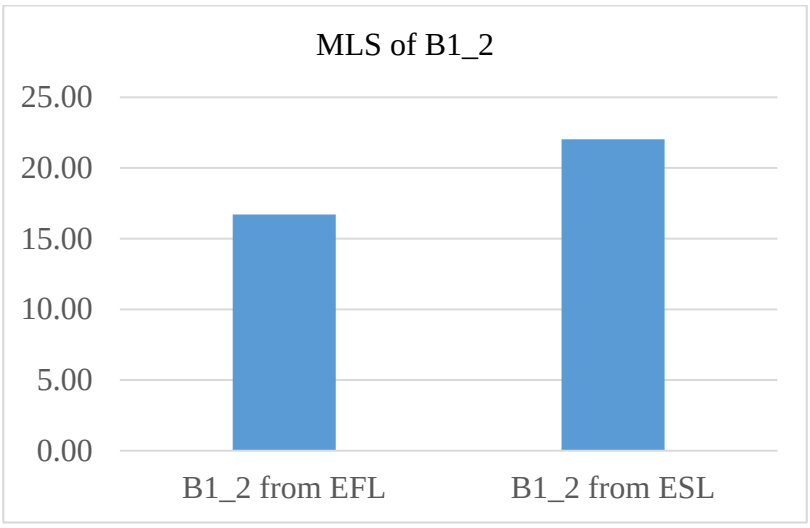
Figure 3-3



Note:
A2_0: (Waystage), B1_1 (Threshold: Lower), B1_2 (Threshold: Upper), B2_0: (Vantage or higher)
MLS: Mean Length of Sentences
Figure 3 MLS of EFL, ESL and ENL

Moreover, despite the shared proficiency level of EFL and ESL in level B1_2, the statistical values for syntactic complexity in terms of mean length of sentences are still statistically different. As illustrated in Figure 4, Both EFL participants and ESL participants with the same proficiency level B1_2 do exhibit quite different levels of syntactic complexity in terms of mean length of sentences. Such a finding further supports the earlier observation that within a certain group, the global complexity measure is

relatively stable, no matter there are some obvious differences of proficiency levels or not. In other words, even though there are some shared proficiency levels between EFL learners and ESL learners, their sentence length is still more related to their language backgrounds rather than their proficiency levels.



Note:
B1_2 (Threshold: Upper)
MLS: Mean Length of Sentences
Figure 4 MLS of proficiency level B1_2 in EFL and ESL

The situation of clauses per sentence is actually quite similar to the trend of mean length of sentences. Again, Figure 5 and Figure 6 prove that the syntactic complexity of EFL, ESL and ENL follows a cline and the two diagrams further confirm the previous observation that in terms of global complexity, language group rather than proficiency level plays a more important role in the differences of syntactic complexity. For learners with the same proficiency level B1_2 from EFL and ESL, the differences of this measure are still quite significant. In addition, ENL writers exhibit much greater variation in this measure with a standard deviation of 0.94 while the figure for EFL and ESL is just around 0.5.

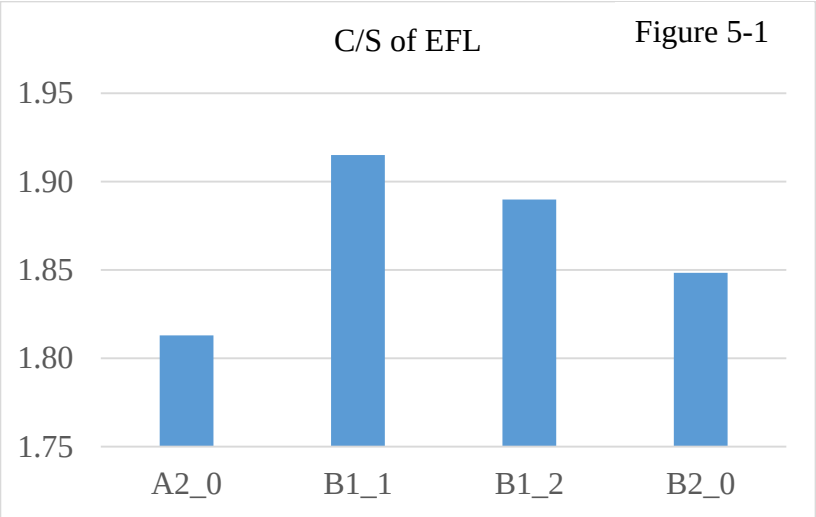


Figure 5-1

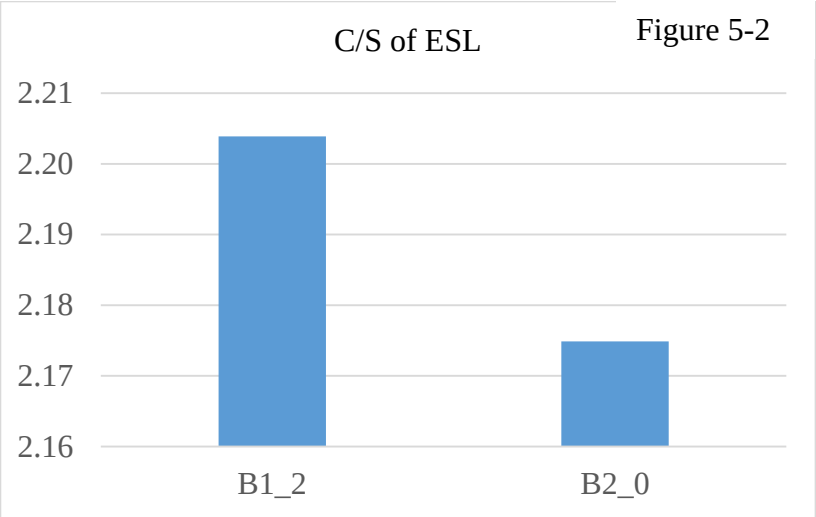


Figure 5-2

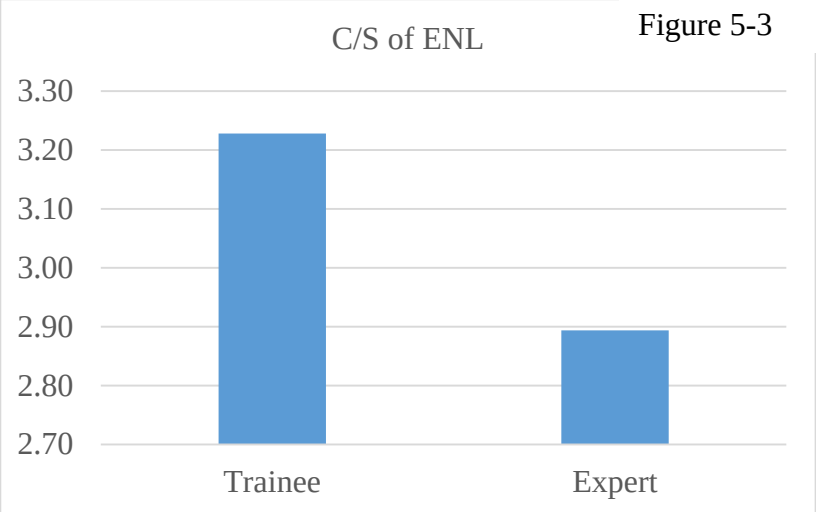
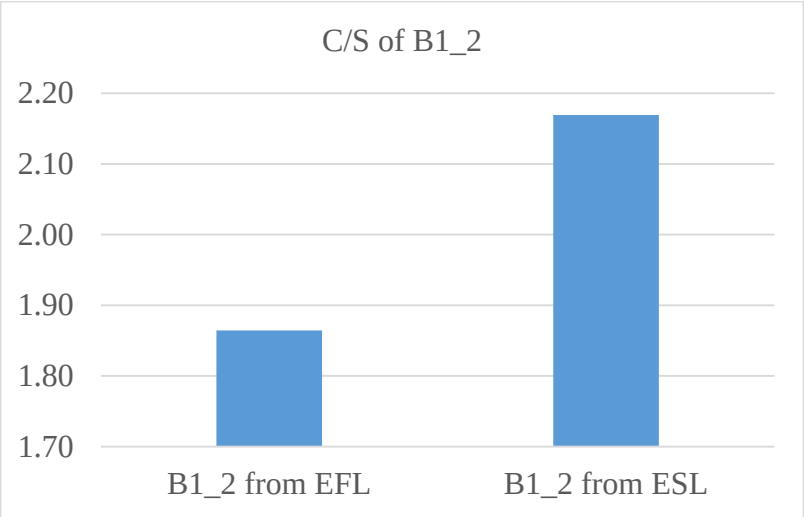


Figure 5-3

Note:
A2_0: (Waystage), B1_1 (Threshold: Lower), B1_2 (Threshold: Upper), B2_0: (Vantage or higher); C/S: Clauses per Sentence
Figure 5 C/S of EFL, ESL and ENL

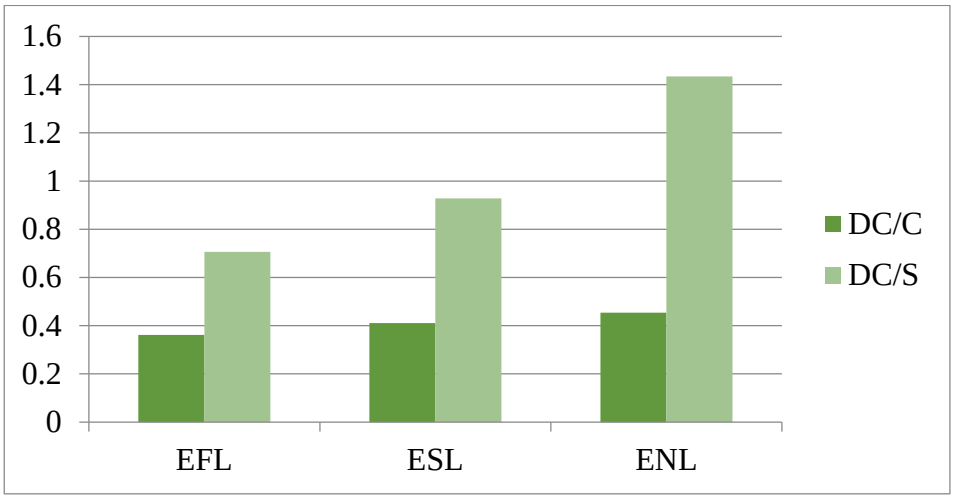


Note:
B1_2 (Threshold: Upper)
C/S: Clauses per Sentence
Figure 6 C/S of proficiency level B1_2 in EFL and ESL

3.2.2 Subordination-based complexity measures and proficiency

Similar to the global syntactic complexity measures, subordination-based complexity measures are also found to be good indicators of proficiency levels.

As shown in Figure 7, both dependent clauses per clause and dependent clauses per sentence do well in signalling different groups across proficiencies. A further examination of the data reveals that compared with number of dependent clauses per clause, number dependent clauses per sentence seems to be a better discriminator for differentiating proficiency levels since the statistics of dependent clauses per sentence from EFL to ENL increases while the statistics of dependent clauses per clause is somehow weaker in signalling the growth of syntactic complexity. According to the statistical analysis, dependent clauses per sentence of ENL is strikingly larger than that of ESL with a figure over 0.5. Dependent clauses per sentence is thus regarded to be a more efficient measure for subordination-based syntactic complexity measure.



Note:
DC/C: Dependent Clauses per Clause; DC/S: Dependent Clauses per Sentence
Figure 7 DC/C and DC/S of EFL, ESL and ENL

Consistent with the observation of global complexity measures, there do not seem to be obvious differences of subordination-based complexity measures within each group despite the identification of proficiencies within them. The three trend lines of Figure 8 illustrate that despite the observable differences of dependent clauses per sentence between each group, no significant differences can be observed in a single group. More specifically, for EFL group, the statistics for dependent clauses per sentence remains around 0.5, regardless of the four proficiency levels. For ESL group and ENL group, the statistics is quite stable although in each group the proficiency levels are identified.

Figure 9 further shows that for participants with the same proficiency level B1_2 from EFL and ESL, the statistics for the subordination-based complexity measure is still quite different, in which ESL group shows obvious higher level of complexity in terms of dependent clauses per sentence compared with EFL group. Such a significant higher level of

syntactic complexity for ESL group is somehow thought-provoking. This may suggest that the association of proficiency with syntactic complexity may not apply to specific proficiency levels within certain groups although based on the research findings, it is quite reasonable to say that each language group is closely related to the certain syntactic complexity level.

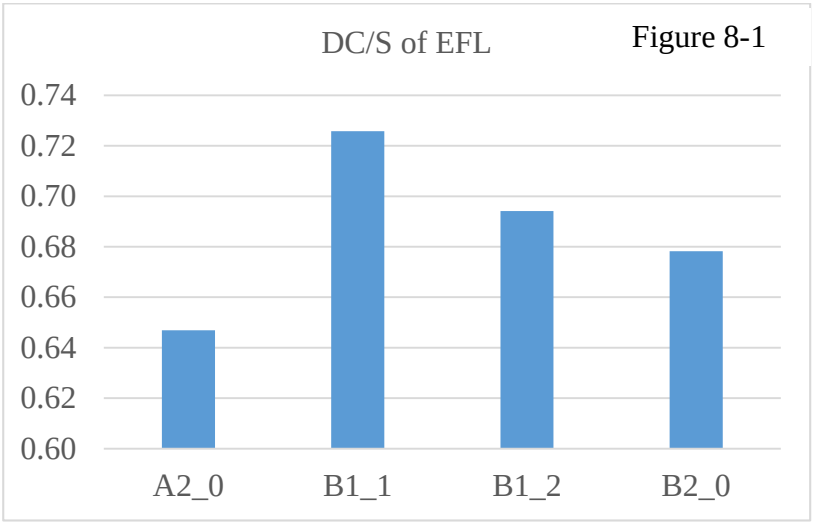


Figure 8-1

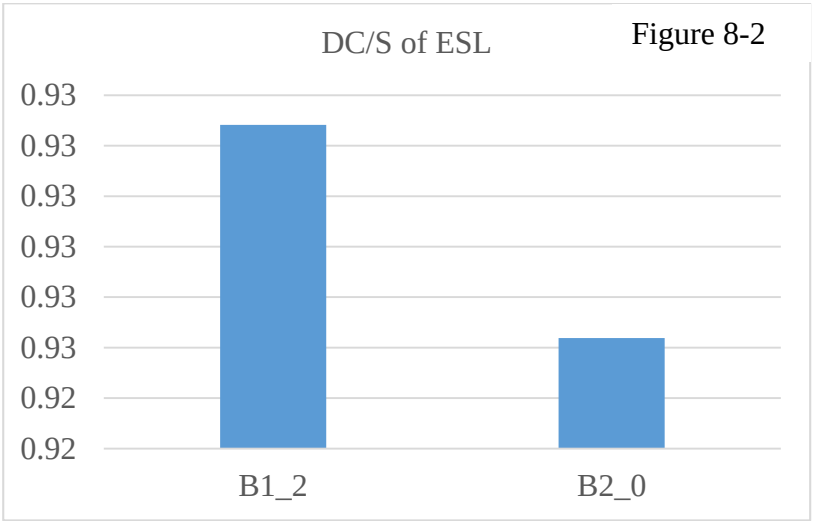
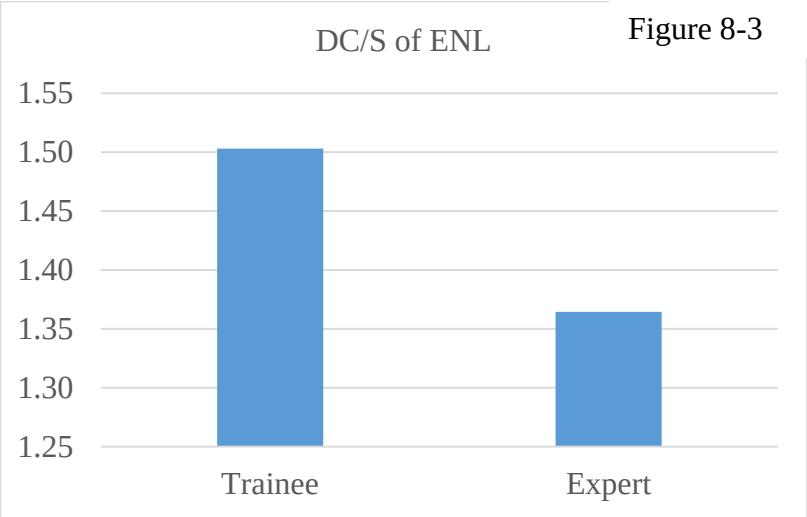
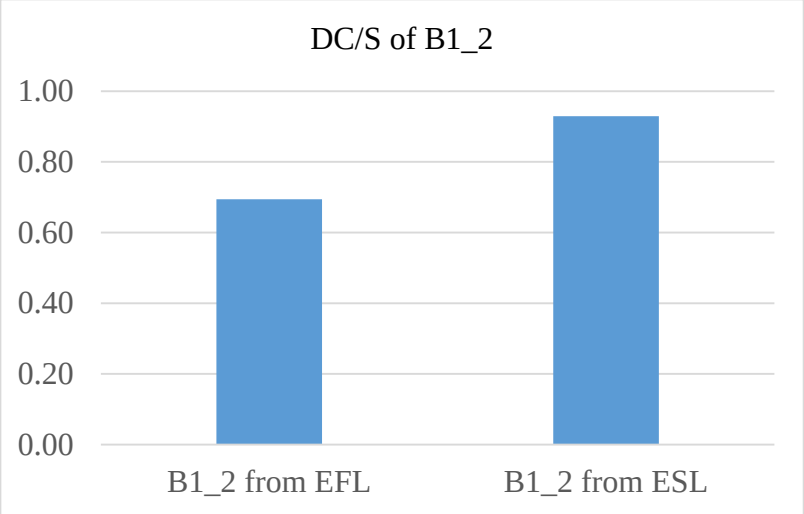


Figure 8-2

Figure 8-3



Note:
A2_0: (Waystage), B1_1 (Threshold: Lower), B1_2 (Threshold: Upper), B2_0: (Vantage or higher)
DC/S: Dependent Clauses per Sentence
Figure 8 DC/S of EFL, ESL and ENL



Note:
B1_2 (Threshold: Upper);
DC/S: Dependent Clauses per Sentence
Figure 9 DC/S of proficiency level B1_2 of EFL and ESL

3.2.3 Coordination-based complexity measures and proficiency

As mentioned earlier, coordination is generally considered to be a typical feature of less advanced technique in sentence complexification. The research findings as shown in Table 7, however, suggest a more complex situation. First, both ESL and ENL, the two more advanced groups use

considerably more coordinate structures compared with EFL learners. Besides, in terms of coordinate phrases per clause, ESL learners are found to use greater number of coordination structures compared with their EFL counterpart and ENL writers. Against the previous expectation, ESL learners rather than EFL learners prefer to use coordination structures in their sentences. Similar to the earlier observation of this research in which measures divided by sentence rather than clause are proved to be more indicative, number of coordinate phrases per sentence seems to be more suitable for discriminating the three groups compared with coordinate phrases per clause. This is especially true in the discrimination of EFL and other two more advanced groups since EFL learners are found to use much less coordinate phrases.

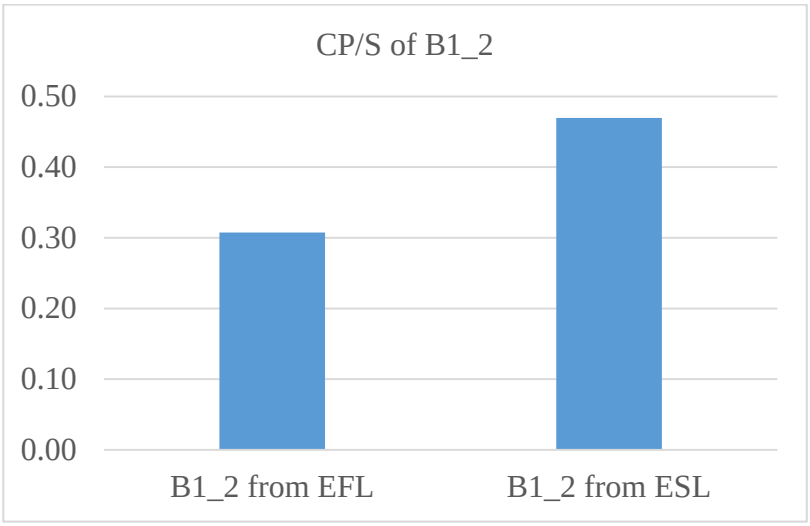
Table 7 Coordination-based complexity measures of EFL, ESL and ENL

Measures	Group	N	Mean	Std. Deviation
CP/C	EFL	800	0.15	0.10
	ESL	400	0.23	0.13
	ENL	400	0.20	0.13
CP/S	EFL	800	0.28	0.18
	ESL	400	0.48	0.28
	ENL	400	0.57	0.31

Note:
CP/C: Coordinate Phrases per Clause: CP/S: Coordinate Phrases per Sentence

Figure 10 compares the complexity measures by coordinate phrases per sentence for learners with the proficiency level of B1_2 in both EFL and ESL, revealing that those EFL learners and ESL learners exhibit quite

different syntactic complexity in terms of number of coordinate phrases per sentence.

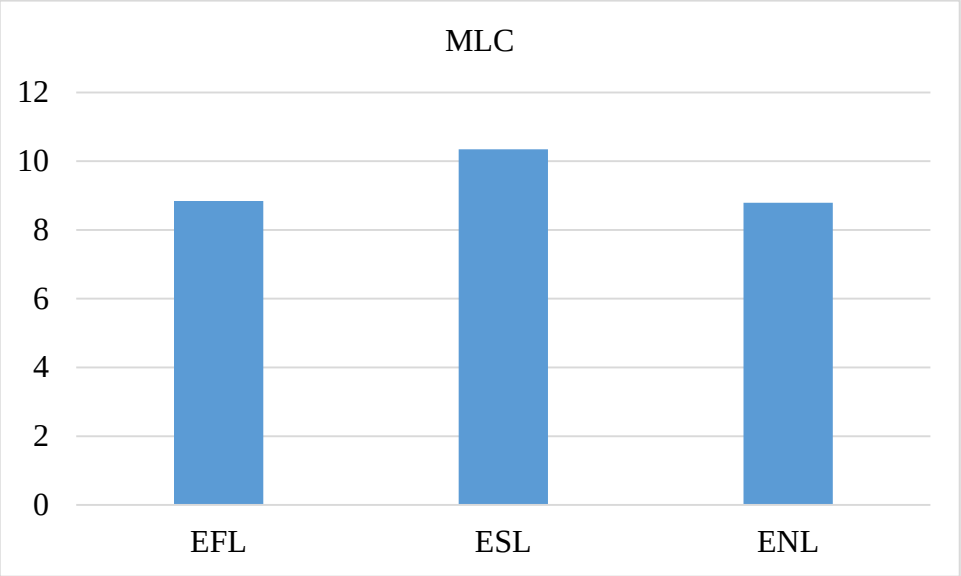


Note:
B1_2 (Threshold: Upper);
CP/S: Coordinate Phrases per Sentence
Figure 10 CP/S of proficiency B1_2 in EFL, ESL and ENL

3.2.4 Phrasal complexity and proficiency

A few linguists have realized the contribution of phrasal complexity to syntactic complexity (e.g. Biber et al., 2011) although phrasal features are not extensively studied in most studies on syntactic complexity. Three measures are involved in the calculation of phrasal complexity of this research while there are several more categories of phrases related to it. The first measure is mean length of clause as generally the use of more complex phrases will increase the length of clauses. Figure 11 has provided a comparison of mean length of clauses in the three groups. With an average length of clauses over 10 words, ESL learners are found to have longer mean length of clauses compared with the EFL learners and ENL writers whose average lengths of clauses are less than 9 words. This discrepancy with the proficiency cline may imply that mean length of clauses is not suitable to

discriminate proficiency levels, which is contradictory to some previous research findings (e.g., Lu, 2011).



Note:
MLC: Mean Length of Clauses
Figure 11 MLC of EFL, ESL and ENL

Among several other categories of phrases, complex nominals are selected to represent phrasal complexity in this research. Complex nominals per clause does not seem to be able to signal the proficiency levels of the three groups while the complex nominals per sentence shows the capability of identifying the differences. Figure 12 indicates the cline of complex nominals per sentences of the three groups. ESL learners and ENL writers who are near the high proficiency end of proficiency cline are found to use more complex nominals (2.54 and 2.90 respectively as shown in Table 8). This is consistent with the anticipation and some previous research findings (e.g., Biber, etal, 2011) that more advanced writing often entails more occurrences of complex nominals.

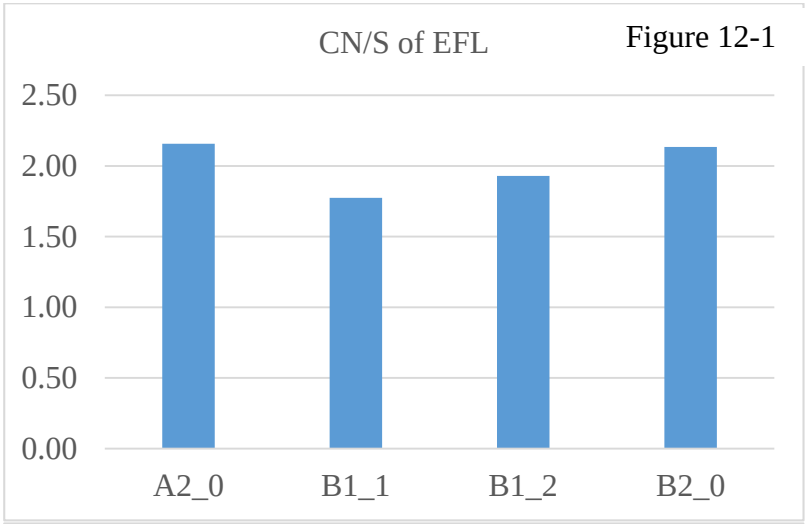


Figure 12-1

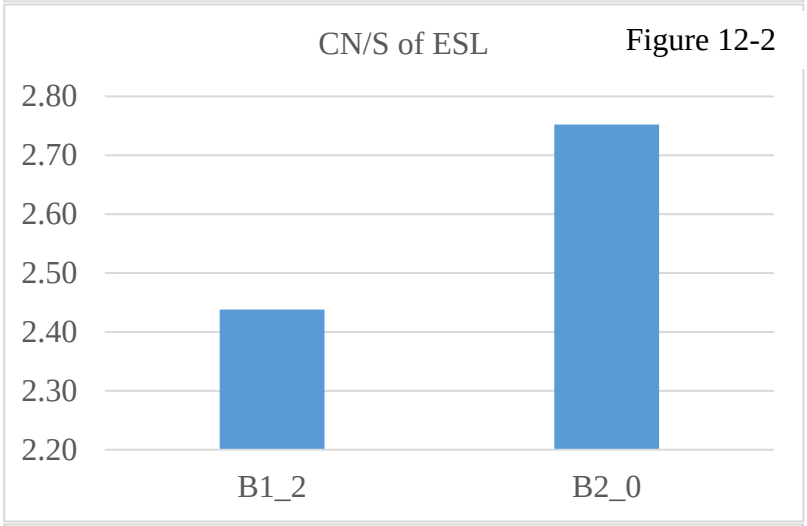


Figure 12-2

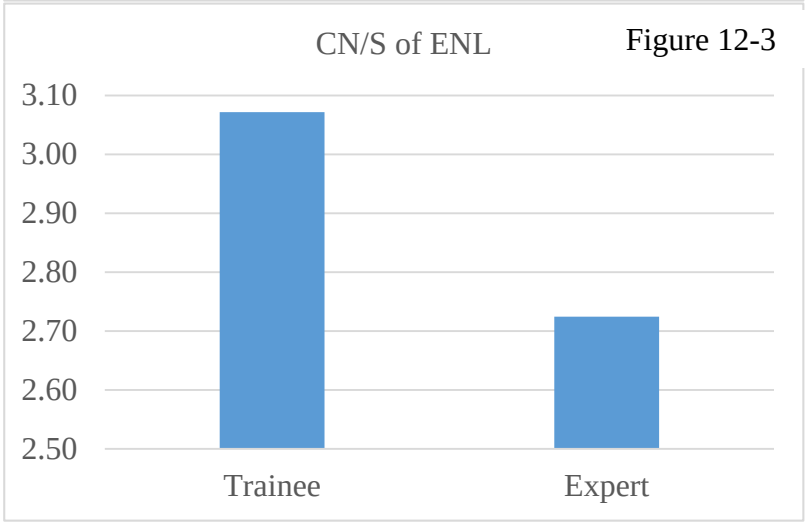


Figure 12-3

Note:
A2_0: (Waystage), B1_1 (Threshold: Lower), B1_2 (Threshold: Upper), B2_0: (Vantage or higher);
CN/S: Complex Nominals per Sentence
Figure 12 CN/S of EFL, ESL and ENL

Table 8 CN/S of EFL, ESL and ENL

Measures	Group	N	Mean	Std. Deviation
CN/S	EFL	800	1.93	0.65
	ESL	400	2.54	0.79
	ENL	400	2.90	0.93

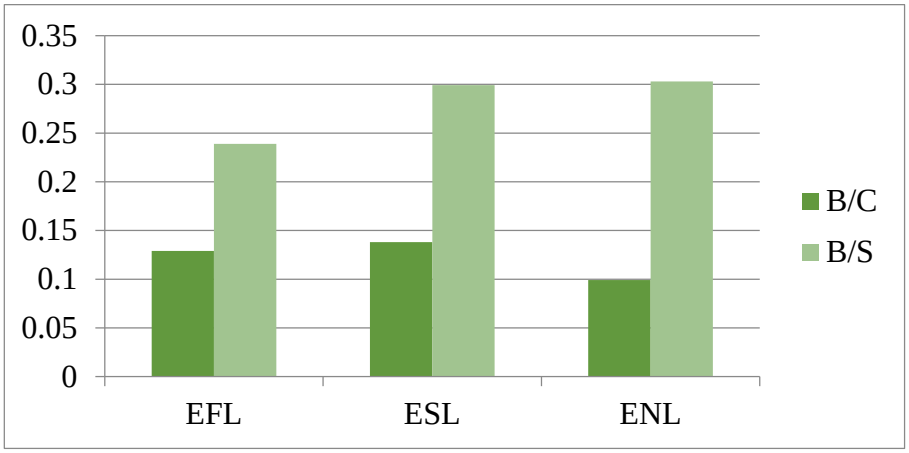
Note:
CN/S: Complex Nominals per Sentence

It is also observed that while the use of complex nominals is relatively stable within EFL and ESL groups despite the differences between proficiency levels, trainee ENL writers seem to use more complex nominals compared with expert ENL writers. This is probably because those expert native writers may use other structures as alternatives of complex nominals in their writing.

3.2.5 Specific complexity measures and proficiency

To further complement the previous measures calculated with automatic annotation tool, four specific complexity measures based on be-copula with adjective structures and it-cleft structures are adopted to uncover some informative insight into the syntactic complexity of the three groups. Figure 13 has illustrated the use of be-copula among the three groups. In comparison with the measure calculated in number of be-copula clauses per clause, number of be-copula clauses divided by number of sentences serves as a better indicator of proficiency level. Surprisingly, contradictory to the previous assumption that be-copula may be overused by EFL learners who are less proficient, ESL learners and ENL writers use much more be-copulas in their writing in terms of the two complexity measures. It is

quite easy to spot that EFL learners actually do not overuse be-copula as expected earlier.



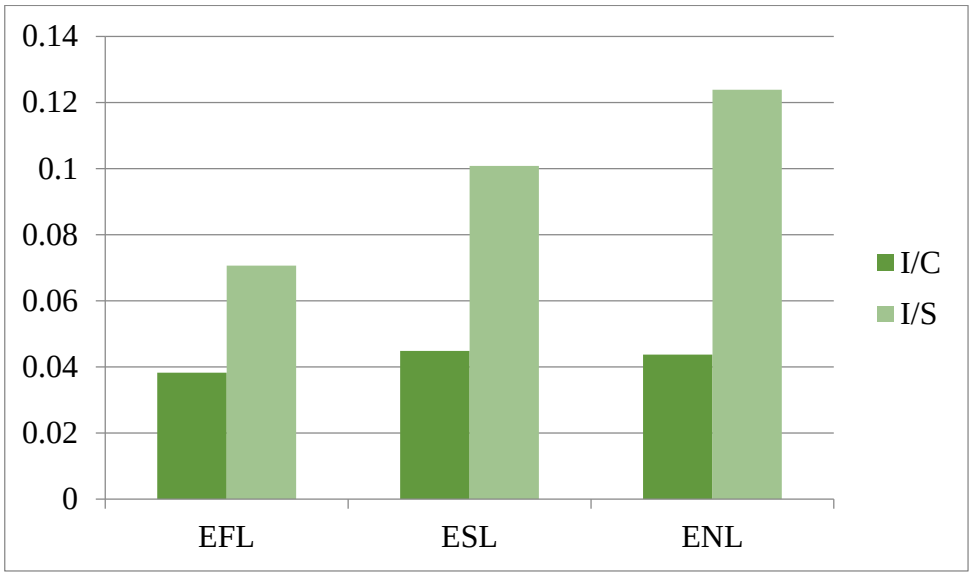
Note:
B/C: Be-copula with Adjective Structures per Clause; B/S: Be-copula with Adjective Structures per Sentence
Figure 13 B/C and B/S in EFL, ESL and ENL

A closer examination indicates that be-copula is actually used by EFL learners with some repetitive expressions like it is (very) / good/ important/ bad/ necessary. For instance, there are 89 occurrences of “is bad” in EFL participants. Figure 14 provides the typical usage of be-copula among EFL learners. On the other hand, apart from the absolute higher ratio, ENL and ESL writers are found to be able to use more varied expression of be-copula in their writing, which is probably because of their larger repertoire of vocabulary. This may suggest another important issue: vocabulary, especially the lexico-grammatical aspects of them, may also play an important role in syntactic complexity because without sophisticated vocabulary, more complex syntactic structures are impossible. As observed in many early studies, vocabulary and syntax are often inseparable.

staurants depends on the real case though smoking is proved to	be	bad	for people's health. And in most of cases smoking in the restauran
vith you, which may lead a conflict. Smoking in public restaurants	is	bad	for people's health so it should be signed that "Please stop smoki
dinner, their behaviors of smoking will not affect others. Though it	is	bad	for their health, it is their choices and they smoke for more enjoym
age their digest system and other organs. In summary, smoking	is	bad	for our health, especially during or after meal. I think the governme
, the air will be polluted. So many people breathe the frog, which	is	bad	for their lungs and will influence people's moods. In my opinion, the
ıte place like sitting room or balcony. Everybody knows smoking	is	bad	for health, including those heavy smokers. The problem is that they
he following are the most obvious aspect. To start with, smoking	is	bad	for people's health. People who smoke frequently more probably h
als better after smoking. However, they didn't know that smoking	is	bad	for their health and to others' health. It can cause lancer, and lead t
ne bans to limit the smokers properly. We all know that smoking	is	bad	for health, not only for the smokers themselves, but also the innoce
what he should learn and have no time to exercise. Obviously, it	is	bad	for his schoolwork and his health. Furthermore, he even doesn't ha
ost people don't like smoking and they always think the smoking	is	bad	for our health. When this type of people goes to the restaurant, they
w cigarettes contain nicotine and other harmful things. Smoking	is	bad	for our health. It has proved that many serious diseases like lung c
r health More and more people have come to know that smoking	is	bad	for people's health. It can cause lung cancer, lead to respiratory pr
ely banned at all the restaurants in the country, because smoking	is	bad	for health, and smoking in restaurants put bad effect to other peop
While smoking, people produce a lot of harmful gas into the air. It	is	bad	to our environment. What's more, the behavior people smoke in th
nd deeply affect the customers around. As we all know, smoking	is	bad	for our health. Nowadays, more and more people who never smok
one smoke there it can affect other people, people' emotion can	be	bad	then so they may not come to the restaurant to have meals any mo
eople. It has not been a simple of fashion but a enemy. Smoking	is	bad	to your health, which I need not say. However, many people agree
e. But I hope that they can control themselves because it will also	be	bad	for other people who do not smoke. Even they do not care about th
little. All the people are then treated equally. To sum up, smoking	is	bad	to some extent, but it's too strict to forbid smoking in public places
Smoking has long been considered to	be	bad	to our health while still there are so many people who can not get ri

Figure 14 Typical use of be-copula by EFL learners

As for the use of it-cleft structures, probably due to the infrequency of it in the three sub-corpora, there is no strong statistical correlation between number of it-cleft structures per clause and proficiency of the three groups observed. Probably a larger database with more occurrences of it-cleft structures can offer more reliable insight into this problem. However, number of it-cleft structures per sentences is found to differentiate the three groups of participants. Similar to what has been observed earlier, measures divided by sentences seem to be better indicators of syntactic complexity compared with those divided by clauses.



Note:
I/C: It-cleft Structures per Clause; I/S: It-cleft Structures per Sentence

Figure 15 I/C and I/S in EFL, ESL and ENL

3.2.6 T-unit-related measures for syntactic complexity

T-unit-related measures, the long established notion for evaluating syntactic complexity is disputable in some recent studies. To further study the feasibility of them, those eight T-unit-related measures produced by the automatic tool L2 Syntactic Complexity Analyser merit a discussion here. The findings of those T-unit-related measures support the latest argument that T-unit-related measures are not quite satisfying in signalling syntactic complexity.

As shown in Table 9, the statistical findings reveal that among the eight measures related to T-units, only verb phrases per T-unit, clauses per T-unit, dependent clauses per T-unit and complex T-units per T-unit are found to be able to discriminate the three groups while the other four could not. The other four measures, however, are able to signify the proficiency levels across the three groups. For instance, when it comes to mean length of T-units, ESL and ENL participants show little difference, failing to

differentiate the two groups. It thus seems quite reasonable to exclude T-unit-related measures in the multidimensional annotation scheme for the current research.

When it comes to mean length of T-units, coordinate phrases per T-unit and complex nominals per T-unit, both ESL group and ENL group seem to be quite similar although the distinction between EFL and these two groups are quite striking. However, in terms of T-units per sentence, both EFL and ESL groups are quite similar while the ENL group shows significantly higher statistical value. Such a complicated situation indicates that those T-unit-related measures are not straight-forward and indicative of proficiency levels.

Table 9 T-Unit-related measures for syntactic complexity

	MLT	VP/T	C/T	DC/T	T/S	CT/T	CP/T	CN/T
EFL	14.96	2.24	1.71	0.64	1.10	0.48	0.25	1.71
ESL	19.95	2.80	1.95	0.83	1.12	0.58	0.43	2.28
ENL	20.09	3.05	2.35	1.11	1.29	0.69	0.46	2.27

Note:
MLT: Mean Length of T-units; VP/T: Verb Phrases per T-unit; C/T: Clauses per T-unit; DC/T: Dependent Clauses per T-unit; T/S: T-unit per Sentence; CT/T: Complex T-unit per T-unit; CP/T: Coordinate Phrases per T-unit; CN/T: Complex Nominals per T-unit

On the whole, there seems to be four important observations in the analysis related to the first research question. First, a strong correlation between global/ subordination-based syntactic complexity measures and language proficiency is observed while the correlation between coordination-based/ phrasal/ specific complexity measures and language proficiency level seems to be dependent on whether sentences or clauses are involved in the calculation. More specifically, all global syntactic

complexity measures and subordination-based measures used in this research seem to be quite useful in discriminating language proficiency levels. It is contradictory to the initial expectation that EFL learners may use more coordinate structures and be-copula structures. Actually, they use both structures much less compared with participants from ESL and ENL. Surprisingly, mean length of clauses are not found to signal proficiency levels as some early studies have found (e.g., Lu, 2011). Second, the data analysis suggests that what differentiates the syntactic complexity is not proficiency alone but also language group. More specifically, a certain group of participants tend to exhibit similar syntactic complexity levels, regardless of their proficiency levels. Learners with the identical proficiency level of B1_2 from both EFL and ESL, for instance, show quite different levels of syntactic complexity. Third, measures divided by sentences rather than clauses are almost always better indicators of proficiency levels. For instance, number of be-copula structures per clause does not signal proficiency in the three groups while number of be-copula structures per sentence does well. Last, compared with EFL learners, ESL learners and ENL writers tend to show more variations in terms of those syntactic complexity features, as suggested by the standard deviation in statistical analysis. Such more observable variations are probably because those more advanced language users (ESL learners and ENL users) may have more options in their language use whereas less proficient EFL learners are generally restricted to a limited number of strategies in writing, resulting in less varied statistics.

3.3 Correlation between syntactic complexity measures

Given the possible links between certain syntactic complexity measures, further correlation analysis is conducted to reveal a clearer picture of syntactic complexity features. Among a few other pairs of correlations, Table 10 to Table 13 offer the correlation values (Pearson's Correlation) of selected measures which merit exploration since those correlation values are relatively high compared with other pairs. Due to the scope of this research, those less observable correlations are excluded from discussion. As for the interpretation of the correlation value, the closer the correlation value is to 1, the more the two measures are positively correlated. On the contrary, -1 signifies an extremely negative correlation between measures.

3.3.1 Subordination-based and global syntactic complexity measures

Table 10 has shown that there is a strong correlation between subordination based measures and global syntactic complexity measures. According to the statistics of Pearson's correlation, the p-values for all correlations are less than 0.00, indicating a strong significance of the result. This is especially true for dependent clauses per sentence and clauses per sentence. It is acceptable to assume that subordination has contributed significantly to global syntactic complexity, resulting in the strong correlational link between dependent clauses per sentence and mean length of sentence/clauses per sentence. The other subordination-based complexity measure, dependent clauses per clause also correlates with global complexity measures positively. In this regard, it is possible to infer that subordination has contributed to the global complexity significantly.

Table 10 Pearson's correlation between subordination-based and general syntactic complexity measures

		MLS	p-value	C/S	p-value
DC/C	Whole	0.54*	0.00	0.54*	0.00
	EFL	0.40*	0.00	0.44*	0.00
	ESL	0.44*	0.00	0.51*	0.00
	ENL	0.47*	0.00	0.46*	0.00
DC/S	Whole	0.79*	0.00	0.91*	0.00
	EFL	0.68*	0.00	0.83*	0.00
	ESL	0.67*	0.00	0.88*	0.00
	ENL	0.74*	0.00	0.89*	0.00

Note:
MLS: Mean Length of Sentences; C/S: Clauses per Sentence; DC/C: Dependent Clauses per Clause; DC/S: Dependent Clauses per Sentence
*. Correlation is significant at the 0.01 level

3.3.2 Coordination-based and global syntactic complexity measures

The correlation between coordination-based measures and global syntactic complexity measures also deserves discussion here. Table 11 illustrates the strong correlation between coordinate phrases per sentence and mean length of sentences. In other words, more frequent use of coordinate phrases may contribute to the length of sentences. It is also noticed that the measure of coordinate phrases per clause, however, does not seem to correlate significantly to global complexity. When it comes to ENL group, however, as the p-value is 0.14, there is no observed statistical significance between coordinate phrases per clause and mean length of sentences, which may suggest that native writers may rely less on coordinate phrases in increasing the length of sentences. Statistics also suggest that for native

writers, there is no tangible correlation between clauses per sentence and coordinate phrases per sentence because of the p-value is 0.84.

Table 11 Pearson's correlation between coordination-based and general syntactic complexity measures

		MLS	p-value	C/S	p-value
CP/C	Whole	0.25*	0.00	-0.13*	0.00
	EFL	0.21*	0.00	-0.18*	0.00
	ESL	0.28*	0.00	-0.19*	0.00
	ENL	-0.07	0.14	-0.46*	0.00
CP/S	Whole	0.60*	0.00	0.32*	0.00
	EFL	0.50*	0.00	0.19*	0.00
	ESL	0.55*	0.00	0.19*	0.00
	ENL	0.33*	0.00	0.01	0.84

Note:
MLS: Mean Length of Sentences; C/S: Clauses per Sentence; CP/C: Coordinate Phrases per Clause; CP/S: Coordinate Phrases per Sentence
*. Correlation is significant at the 0.01 level

3.3.3 Phrasal, global and subordination-based complexity measures

Closer examination of the statistics reveals that there are also important correlations between phrasal, global and subordination-based complexity measures. As shown in Table 12 mean length of clauses is found to be negatively correlated to clauses per sentence and dependent clauses per sentence. This is quite understandable because generally the longer the clause is, the less clauses per sentence will be. Besides, a longer clause may often involves longer independent clauses as modifiers, as a result, the dependent clauses become relatively shorter and the value of dependent clauses per sentence also decreases.

Besides, complex nominals per sentence is found to be strongly related to both mean length of sentences and clauses per sentence, suggesting the contribution of complex nominals to sentence length and the ratio between clauses and sentences. Complex nominals per sentence is also found to influence the occurrence of dependent clauses per sentence, given the high value of statistical correlation. This is probably because in many occasions dependent clauses may constitute complex nominals. Again, it is noted that the measures of complex nominals per clauses does not show strong correlations with other measure, supporting the use of measures divided by sentences. Moreover, for native writer group, no statistical significance (p-value: 0.14) can be established when it comes to the correlation between mean length of clauses and mean length of sentences. This is probably because native writers may have more varied writing techniques and preferences compared with the other two learners' groups. Similarly, for native writers, number of complex nominals per clause and mean length of sentences are not correlated (p-value: 0.78) based on statistical examination. In this regard, it seems it is more difficult to infer native writers' complexification strategies compared with other two groups of learners.

Table 12 Pearson's correlation between phrasal and global/ subordination-based syntactic complexity measures

		MLS	p-value	C/S	p-value	DC/S	p-value
MLC	Whole	0.18	0.00*	-0.40	0.00*	-0.32	0.00*
	EFL	0.28	0.00*	-0.40	0.00*	-0.30	0.00*
	ESL	0.31	0.00*	-0.44	0.00*	-0.32	0.00*
	ENL	-0.07	0.14	-0.66	0.00*	-0.54	0.00*
CN/C	Whole	0.17	0.00*	-0.29	0.00*	-0.20	0.00*
	EFL	0.29	0.00*	-0.24	0.00*	-0.13	0.00*
	ESL	0.28	0.00*	-0.31	0.00*	-0.20	0.00*
	ENL	-0.01	0.78	-0.50	0.00*	-0.39	0.00*
CN/S	Whole	0.81	0.00*	0.59	0.00*	0.60	0.00*
	EFL	0.70	0.00*	0.44	0.00*	0.44	0.00*
	ESL	0.79	0.00*	0.45	0.00*	0.47	0.00*
	ENL	0.79	0.00*	0.49	0.00*	0.53	0.00*

Note:
MLS: Mean Length of Sentences; C/S: Clauses per Sentence; DC/S: Dependent Clauses per Sentence; MLC: Mean Length of Clauses; CN/C: Complex Nominals per Clause; CN/S: Complex Nominals per Sentence
*. Correlation is significant at the 0.01 level

3.3.4 Measures related to mean length of clauses

Clauses, as the first degree component of sentences, are also influenced by many other structures. Statistical results illustrated on Table 13 also indicate that the mean length of clauses is positively associated with two measures, namely, coordinate phrases per sentence and complex nominals per clause (all p-values are smaller than 0.001). It is not difficult to infer that coordinate phrases and complex nominals can contribute to the length of

clause. Both of them are important techniques for increasing the length of clauses.

Table 13 Pearson's correlation between MLC and other measures

		CP/C	p-value	CN/C	p-value
MLC	Whole	0.62*	0.00	0.80*	0.00
	EFL	0.59*	0.00	0.76*	0.00
	ESL	0.61*	0.00	0.77*	0.00
	ENL	0.66*	0.00	0.84*	0.00

Note:
MLC: Mean Length of Clauses; CP/C: Coordinate Phrases per Clauses;
CN/C: Complex Nominals per Clause
*. Correlation is significant at the 0.01 level

On the whole, there are primarily four groups of strong correlations between those measures. First, global complexity and subordination-based complexity measures are strongly correlated with each other. Second, number of coordinate phrases per sentence is strongly correlated to global syntactic complexity while coordinate phrases per clause does not. Third, mean length of clauses and clauses per sentence/ dependent clauses per sentence are negatively correlated with each other while complex nominals per sentence rather than per clause is also strongly correlated with clauses per sentence/ dependent clauses per sentence. Last, coordinate phrases and complex nominals are found to be strongly related to the length of clauses, probably because in many occasions complex nominals and coordinate phrases are important sources in increasing the length of clauses.

3.4 Effect of topic on syntactic complexity

Because of the strict control of topics in the ICNALE Corpus, the comparison of topic effect is feasible in this research. All the measures of

syntactic complexity are thus further analysed according to the two topics. Both the effects of topic on the three groups as a whole and each group individually are discussed here. Table 13 provides an overview of the influence of topic on the three groups, covering the statistical values for both topics in line with the 13 measures used in this research.

3.4.1 General comparison of syntactic complexity in two topics

Before moving on to the influence of topic on certain category of complexity measures, a quick glance of the statistics also reveals some interesting information. It seems that on the whole, there are obvious differences of syntactic complexity for the two topics, as is shown in Table 14. Topic on part-time job seems to induce higher syntactic complexity in terms of most syntactic complex measures adopted in this research, based on the higher statistics of topic part-time job over smoking in the majority of measures as shown in Table 14.

Overall, the topic effect applies to the mean length of sentences, coordination-based complexity measures, phrasal complexity measures and measures related to be-copula with adjective structures. In other words, among all those 13 measures, it is found that the majority, or 9 of those measures are subject to the influence of topic change. This is a strong support that certain topics can induce more complex syntactic structures compared with others. More specifically, certain topics may have their advantages in soliciting more syntactically complex sentences, for instance, longer sentences and more coordinate structures.

The two subordination-based measures and two specific measures related to it-cleft structures, however, are not strongly influenced by topic.

Values for subordination-based measures for the two groups do not follow a certain cline. In addition, the insensitivity of it-cleft structures to topic effect as shown in the statistical analysis is primarily because its infrequency.

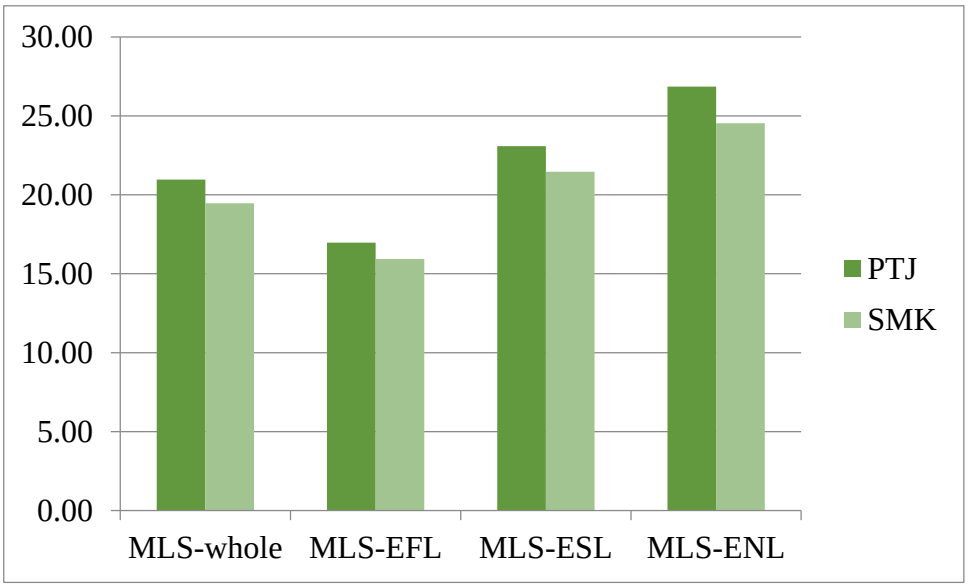
Table 14 Topic effect on the whole data and each group

		MLS	C/S	DC/C	DC/S	CP/C	CP/S	MLC	CN/C	CN/S	B/C	B/S	I/C	I/S
WHOLE	PTJ	20.97	2.25	0.40	0.95	0.20	0.44	9.61	1.11	2.42	0.14	0.29	0.04	0.09
	SMK	19.47	2.27	0.40	0.94	0.16	0.36	8.80	0.99	2.19	0.11	0.25	0.04	0.10
EFL	PTJ	16.97	1.87	0.36	0.70	0.17	0.31	9.23	1.08	1.99	0.14	0.25	0.04	0.07
	SMK	15.94	1.92	0.36	0.72	0.13	0.25	8.46	0.94	1.77	0.12	0.23	0.04	0.07
ESL	PTJ	23.08	2.18	0.41	0.93	0.26	0.56	10.79	1.23	2.62	0.16	0.35	0.05	0.11
	SMK	21.46	2.21	0.41	0.93	0.19	0.40	9.90	1.14	2.46	0.11	0.25	0.04	0.09
ENL	PTJ	26.85	3.08	0.47	1.48	0.21	0.59	9.19	1.05	3.05	0.11	0.32	0.04	0.13
	SMK	24.54	3.04	0.44	1.39	0.20	0.55	8.40	0.95	2.74	0.09	0.28	0.05	0.12

Note:
MLS: Mean Length of Sentences; C/S: Clauses per Sentence; DC/C: Dependent Clauses per Clause; DC/S: Dependent Clauses per Sentence; CP/C: Coordinate Phrases per Clauses; CP/S: Coordinate Phrases per Sentence; MLC: Mean Length of Clauses; CN/C: Complex Nominals per Clause; CN/S: Complex Nominals per Sentence; B/C: Be-copula with Adjective Structures per Clause; B/S: Be-copula with Adjective Structures per Sentence; I/C: It-cleft Structures per Clause; I/S: It-cleft Structures per Sentence

3.4.2 Influence of topic on mean length of sentences

Figure 16 has shown that obviously all three groups of participants produced longer sentences for the topic on part-time job. Statistics indicates that for the topic on part-time job, the average length of sentences by EFL learners is 1.03 words longer than the length for topic on smoking. For ESL learners, there are 1.62 words longer while for ENL writers there are even 4.85 words longer. It seems to suggest that the sentence length of ESL learners and ENL writers, the more advanced groups, is actually more sensitive to the topic.



Note:
MLS: Mean Length of Sentences

Figure 16 Topic effect on mean length of sentences

3.4.3 Influence of topic on subordination and coordination

The use of subordination seems to be uncertain for the two topics while the use of coordination is found to be influenced by the topic effect. Such effect is especially obvious for both EFL and ESL groups while ENL writers are less influenced by topic in terms of coordination-based complexity measures. A closer examination may reveal some interesting

observations about the topic impact of subordination. Figure 17 has shown that for ENL writers, there are still some noticeable differences of subordination for the two topics.

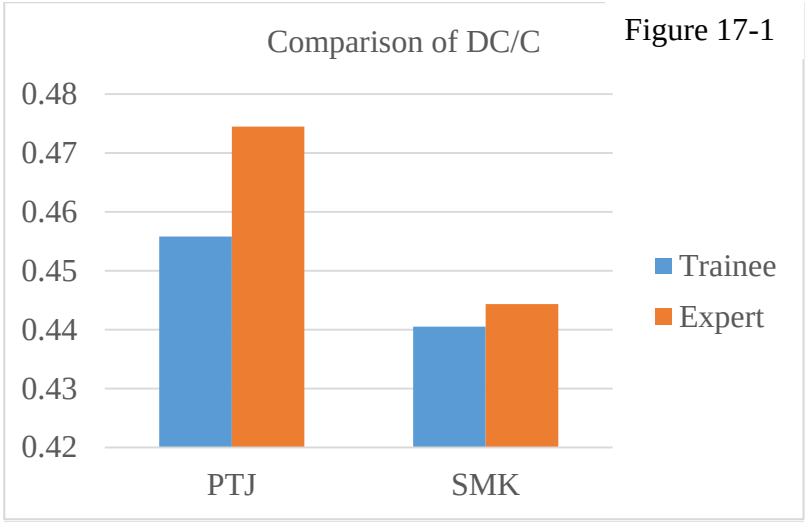


Figure 17-1

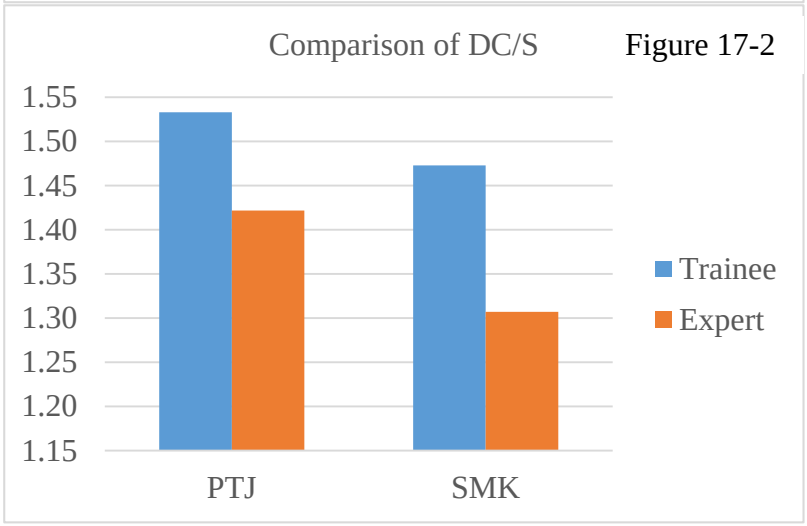
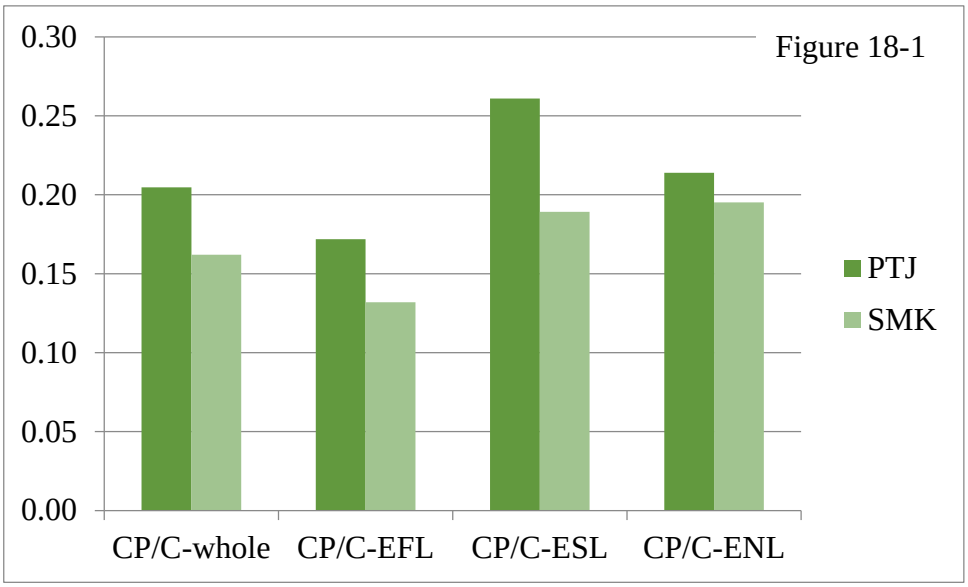


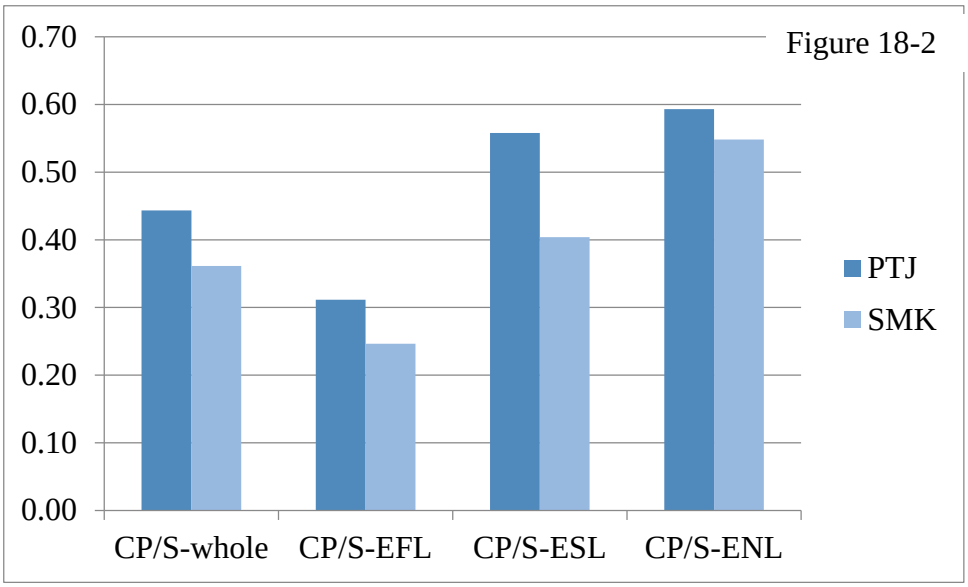
Figure 17-2

Note:
DC/C: Dependent Clauses per Clause; DC/S: Dependent Clauses per Sentence
Figure 17 Topic effect on subordination by ENL

Nevertheless, the use of coordination is obviously different in two topics across the three groups. A closer examination suggests that EFL learner and ESL learner, compared with ENL writers, are influenced to a larger extent by different topics as both of them exhibit higher level of coordination with the topic on part-time job. The use of coordination by ENL writers is found to be less sensitive to topic change since for both topics

there are no striking differences compared with the obvious differences in the two learner groups. Figure 18 presents the effect of topic on coordination-based measures among the three groups.

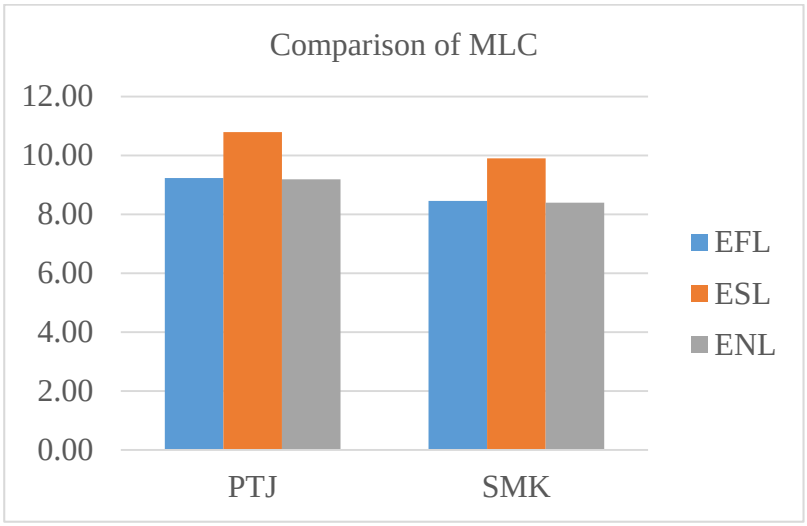




Note:
CP/C: Coordinate Phrases per Clauses; CP/S: Coordinate Phrases per Sentence
Figure 18 Topic effect on coordination

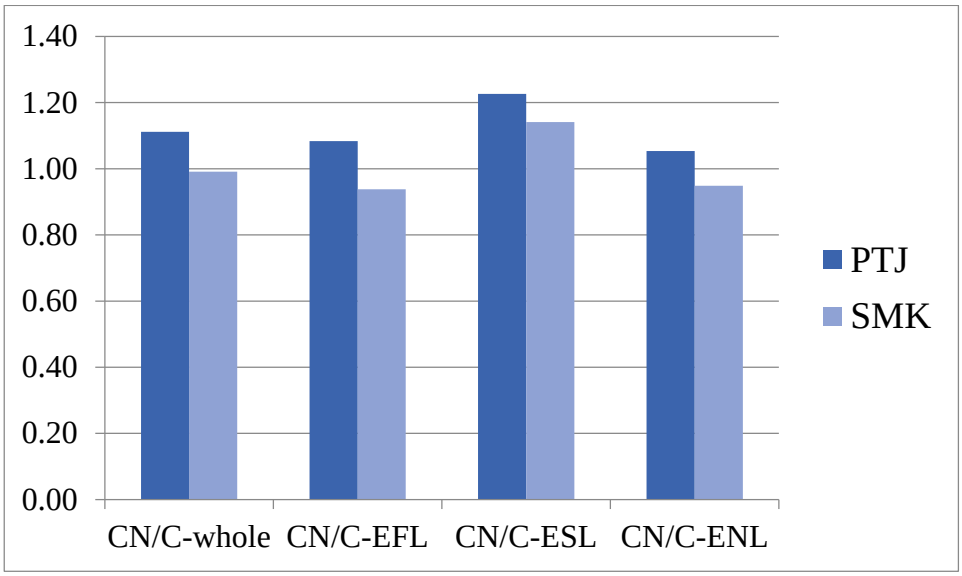
3.4.4 Impact of topic on phrasal complexity

Apart from the observation of its influence on sentence length and coordination, topic is also identified to be associated with phrasal complexity. This influence applies to all the three measures of phrasal complexity. First of all, mean length of clauses is influenced significantly with the topic effect. As figure 19 suggests, all the three groups are found to produce longer average length of clauses with the topic on part-time job.

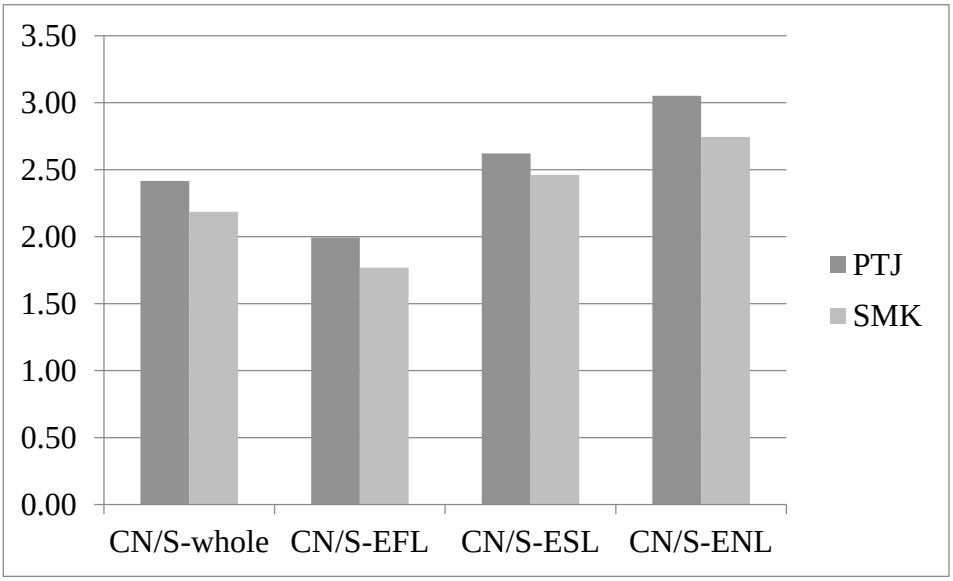


Note:
MLC: Mean Length of Clauses
Figure 19 Topic effect on MLC

Meanwhile, the topic also has an effect on the use of complex nominals. Part-time job seems to afford more use of complex nominals. As researchers increasingly realize the contribution of complex nominals to syntactic complexity, the use of complex nominals in the three groups merits exploration here. Figure 20 and Figure 21 offers an illustration of the topic influence of complex nominals on the three groups.



Note:
CN/C: Complex Nominals per Clause
Figure 20 Topic effect on CN/C



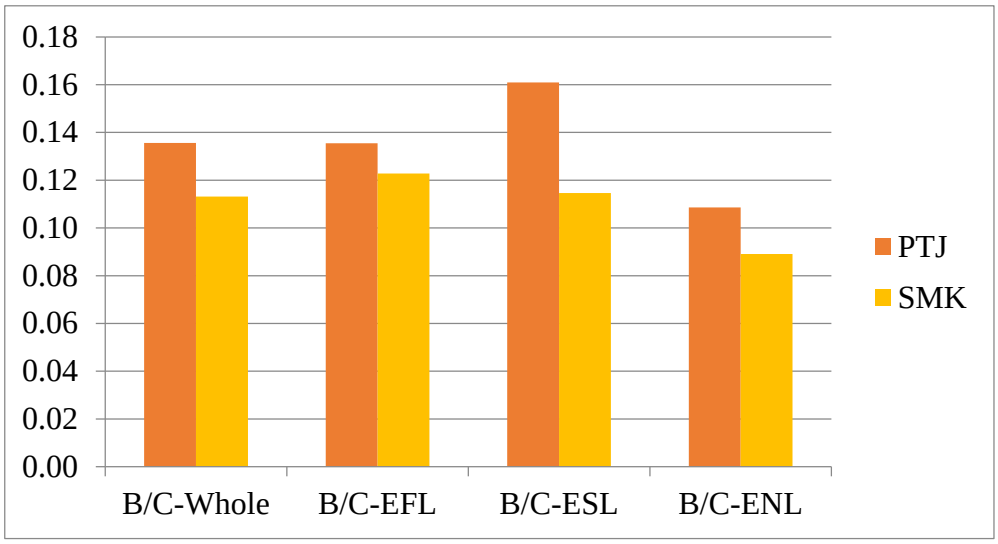
Note:
CN/S: Complex Nominals per Sentence
Figure 21 Topic effect on CN/S

3.4.5 Influence of topic on specific complexity measures

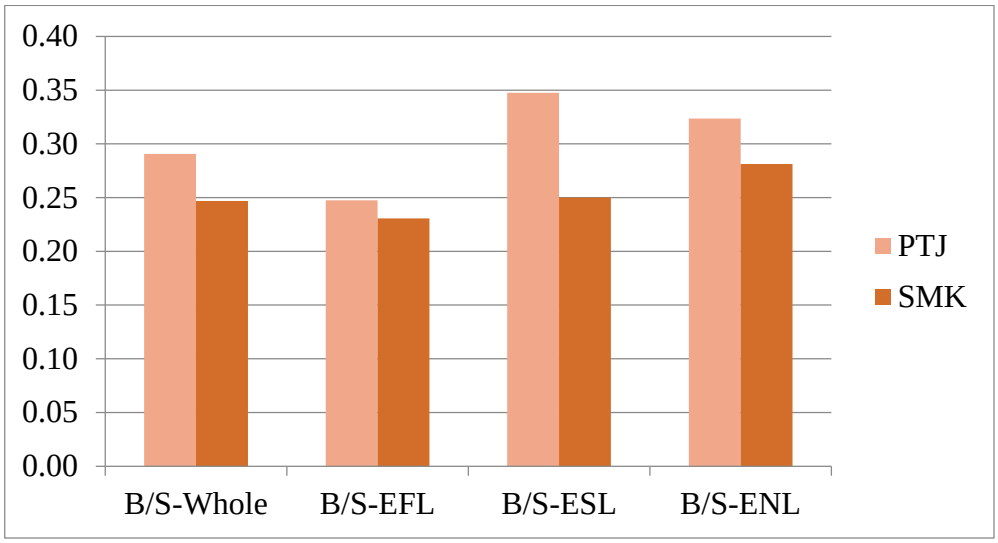
The research findings also reveal the influence of topic on specific complexity measures. It applies to be-copula with adjective structures, including both be-copula with adjective structures per clause and be-copula with adjective structures per sentence. As shown in Figure 22 and 23, for this

pair of measures, topic on part-time job seems to induce higher complexity compared with the topic on smoking. A closer look further indicates that ESL learners seem to be more sensitive to the change of topics with regards to this two measures. EFL learners and ENL writers, however, are relatively insensitive to the topic change.

Another pair of specific complexity measures concentrate on it-cleft structures. However, largely due to the infrequency of such structures in all the three groups, there does not seem to be any observable impact of topic in the three groups in terms of the two syntactic complexity measures.



Note:
B/C: Be-copula with Adjective Structures per Clause
Figure 22 Topic effect on B/C



Note:
B/S: Be-copula with Adjective Structures per Sentence
Figure 23 Topic effect on B/S

On the whole, the topic influence on syntactic complexity is well-supported by the research findings. Topic on part-time job seems to induce higher syntactic complexity when it comes to global syntactic complexity, coordination-based complexity measures, phrasal complexity measures and specific measures based on be-copula with adjective structures. Subordination-based measures and the specific complexity measure based on it-cleft structures are not found to be consistently influenced by topic.

3.5 Chapter Conclusion

This chapter involves detailed analyses of the research data in order to address the three research questions. It is possible to conclude that certain syntactic complexities are considered to be good indicators of proficiency levels. The correlation between certain syntactic complexity measures has also been established. Moreover, topic effect on certain syntactic complexity measures has been identified with a large body of evidence, supporting the

necessity of considering the topic as an important factor of syntactic complexity.

CHAPTER FOUR: DATA DISCUSSION

4.1 Introduction

There are some thought-provoking observations revealed in the data analysis, providing satisfying answers to the research questions. In what follows, the analysis for each research question will be further discussed in an attempt to explain the key findings of this research. Findings from previous studies are compared when necessary. The possible causes of the discrepancies are explained tentatively also, followed by the recommendations for improvement in teaching or pedagogy.

4.2 Syntactic complexity and proficiency

The first research question deals with the relationship between syntactic complexity and proficiency. Research findings highlight that certain syntactic complexity measures are positive indicators of proficiency and others are relatively weak in identifying proficiency levels. On the whole, global complexity measures and subordination-based measures are always positive indicators of proficiency whereas the other four categories of complexity measures fall into positive indicators and weak indicators in identifying proficiency levels. The methodological implications drawn from the data analysis are also discussed to benefit future research, followed by the possible implications for teaching.

4.2.1 Measures serving as positive indicators of proficiency

In addition to global complexity measures and subordination-based complexity measures, measures divided by sentences in the coordination-based/phrasal/specific complexity categories and half of the

eight T-unit-based measures are also found to be positive indicators of proficiency.

4.2.1.1 Global complexity measures

Both of the global complexity measures are proved to be strong indicators of syntactic complexity. The research findings confirm there are significant differences between the three groups in their mean sentence length and number of clauses per sentence, indicating a strong increase of syntactic complexity from EFL to ENL in accordance with the two global syntactic complexity measures. This is especially true for the mean length of sentences. Consistent with many previous findings (e.g., Lu, 2011; Ortega, 2003; Vaezi and Kafshgar, 2012), mean length of sentences is found to be a very useful syntactic complexity measure in differentiating proficiency levels.

The varying average sentence length between the three groups can be explained by further referring to the other syntactic measures like coordination-based complexity measures and coordination-based complexity measures. Similarly, ESL learners and ENL writer show high figures in terms of those complexity measures. As noted by Vyatkina (2012), sentence length can be increased by adding more coordinate or subordinate clauses to a matrix clause (clauses/sentences). The fact that ESL learners and ENL writers have longer average sentences can be accountable to the increased use of subordination and coordination, which is discussed in the following sections.

4.2.1.2 Subordination-based measures

The research findings on subordination-based measures confirm the previous research findings that those subordination-based measures, be they dependent clauses per clause or dependent clauses per sentence, do signal the differences between EFL and ESL (e.g., Ortega, 2003) as well as differences between learners with varying proficiency levels (e.g., Vaezi and Kafshgar, 2012). This can be further extended to differentiate EFL/ESL and ENL. Use of dependent clauses as one of the most important types of syntactic complexity (e.g., Carter & McCarthy, 2006, p. 489; Purpura, 2004, p. 91; Willis, 2003, p. 192) is thus proved to be another ideal indicator of proficiency. Based on the research data, it is necessary to highlight that number of dependent clauses per sentence seems to be a better indicator compared with number of dependent clauses per clause. Such a slight difference between those measures divided by clauses and sentences seems to be quite consistent across different categories of measures.

4.2.1.3 Other categories of measures divided by sentences

Surprisingly, for the other three categories, measures divided by sentences are always found to be able to discriminate proficiency levels while those divided by clauses fail to do so. This applies to coordination-based measures, phrasal complexity measures and specific complexity measures.

Whether coordination should be adopted as category of syntactic complexity measures is also disputable because most previous studies do not include it often and the existing studies tend to regard it as a simple feature of syntactic complexity. For instance, Bardovi-Harlig (1992) argues that “the

measurement of increased clausal complexification achieved via coordination is quite relevant for data at initial levels of L2 development". Bearing such assumption in mind, before analysing the data I think that coordination structures would be overused by EFL learners. However, the data analysis provides a quite different picture. Strangely, both of the two more advanced groups ESL and ENL are found to use considerably more coordinate structures compared with EFL learners. Moreover, the two coordination-based measures show quite different situations when dealing with proficiency. ESL learners are found to exhibit greater number of coordinate phrases per clauses, compared with their EFL counterpart and ENL writers. The other coordination-based measure, number of coordinate phrases per sentence, however, seems to be more suitable for discriminating the three groups because it can match the cline of proficiency of the three groups.

The reason why students at higher proficiency levels tend to use more coordinate phrases per sentence is probably because this is a strategy for them to produce longer sentences while those less proficiency learners do not think too much of it. The research finding in coordination seems to echo the research conducted by Cooper (1976), who noticed that coordinate phrases, among several other measures, increased linearly from lower level to high level. This confirms that while subordination is quite straightforward in signalling proficiency, complexification strategies other than subordination can also be important resources for writers in enhancing complexity (Ortega, 2012). Notably, as mentioned in the discussion of

subordination-based measures, coordination-based measures divided by sentence seems to be more indicative of proficiency levels.

Apart from the distinction between two coordination-based measures, the other two categories of measures also follow the same distinction pattern. Unlike number of complex nominals per clause, number of complex nominals per sentence seems to be an acceptable indicator of proficiency levels. I speculate that the use of complex nominals would not always formulate clauses. As a result of mathematical calculation, the use of it may be less relevant to the measures divided by clause. Instead, sentence-based measures can be closer to its trend. In a similar vein, the use of be-copula and it-cleft follows the similar distinction pattern of coordination-based measures and phrasal complexity measures: the two specific measures divided by sentences are better indicators of proficiency.

4.2.1.4 T-unit-based measures indicative of syntactic complexity

The statistical findings has testified that only half of the eight T-unit-based measures are indicative of proficiency levels, namely, verb phrases per T-unit, clauses per T-unit, dependent clauses per T-unit and complex T-units per T-unit. It is true that the four T-unit-based measures are indicative of proficiency levels. The major problem is that the generalization or classification of those measures is quite difficult since there is no clear clue to the use of them. In this regards, it is quite reasonable to reconsider the use of T-unit-based measures in syntactic complexity research.

4.2.2 Measures serving as weak indicators of proficiency

In addition to the use of mean length of clauses as a syntactic complexity measures, those measures divided by clauses are also found to be

weak indicators of proficiency. As for the T-unit-based measures, the situation is quite complicated.

4.2.2.1 Mean length of clauses as a weak indicator of proficiency

Mean length of clauses as an indicator of syntactic complexity for differentiating proficiency levels seems to be challenged in this research. In other words, the empirical data in this research supports that the average length of clauses does not really differentiate syntactic complexity since in data of this research, ESL participants rather than ENL participants are found to exhibit the longest average length of clauses. Obviously, the research findings indicate that ESL group is found to have a significantly longer mean length of clauses compared with those of EFL group and ENL group. The research findings on mean length of clauses echoes a recent research conducted by Vyatkina (2012) who argued that "the clause-type unit length in words did not work when differentiating proficiency levels". However, as early as several decades ago, Hunt (1970) has already argued that number of words per clause is "one of the three most reliable indicators of syntactic complexity". Some other studies (e.g., Byrnes, 2009; Lu, 2011; Ortega, 2003) also favour that the significant growth of clause length may translate into the increase of proficiency. An important difference between the current research and theirs is that in those studies only ELF or ESL group alone is considered while in the current study EFL, ESL and ENL are all included to make comparisons. Probably it is suitable to say that clause length can be used to discriminate EFL and ESL learners, but it is not necessarily a good indicator for differentiating ESL and ENL groups, or the three groups as a whole. Besides, we also need to note that there is a significant growth of

subordination-based measures for ENL writers compared with ESL learners. In other words, ENL writers may choose to use more embedding rather than longer clauses, resulting in shorter production of clauses. This is consistent with some earlier findings (e.g., Arthur, 1979; Kern & Schultz, 1992).

4.2.2.2 Other categories of measures divided by clauses

For the other three categories of syntactic complexity measures, those measures divided by clauses seem to be weak indicators of proficiency levels. First of all, the use of coordination-based measures divided by clauses is not indicative of proficiency across the three groups. ESL learners are found to exhibit greater number of coordinate phrases per clauses, compared with their EFL counterpart and ENL writers. This may suggest that ESL learner prefer to use more coordinate phrases in their clauses while the other two groups may use coordination less in clauses.

In terms of phrasal complexity measures, ratio of complex nominals per clause does not seem to be able to signal the proficiency levels of the three groups while the ratio of complex nominals per sentence shows the capability of identifying the differences. This is against the observation of Lu (2010), who found that number of complex nominals per clause is "a good indicator of proficiency levels". An important distinction between this research and his is that the current research also includes ESL and ENL data. This is most likely that number of complex nominals per clause is not capable of differentiating the three groups in a cline although it is possible to signal the proficiency levels within EFL learners.

As for specific complexity measures, the use of be-copula and it-cleft structures divided by clauses also does not provide good correlation between

proficiency levels. It is possible because the two structures often constitute single sentences themselves rather than adding number to clauses alone. Consequently, they do not seem to be closed related to measures divided by clauses.

4.3.2.3 T-unit-based measures as weak indicators of syntactic complexity

The situation of T-unit-based measure is quite difficult to generalize: coordinate phrases per T-unit and complex nominals per T-unit are unable to differentiate ESL group and ENL group although they seem to be able to differentiate EFL and these two groups. Moreover, T-units per sentence fails to differentiate EFL and ESL groups while the ENL group shows significantly higher statistical value. All of them do not support the idea that the use of T-unit-based measures are indicative of proficiency levels.

4.2.3 Methodological implications

The distribution of the analysis data seems to shed some lights on the methodological issues: language group rather proficiency seems to impact more on syntactic complexity; measures divided by sentences are found to be more indicative than those divided by clauses; advanced participants, including ESL learners and ENL writers, tend to show more variation in terms of those syntactic complexity measures; T-unit-based measures are somehow difficult to be generalized or categorized for application in syntactic complexity research.

4.2.3.1 Impact of language group of syntactic complexity

Language group rather than proficiency alone may play a key role in differentiating the syntactic complexity, as tested in the comparison of several syntactic complexity measures by B1_2 students from both EFL and

ESL backgrounds. For instance, there are significant differences of coordinate phrases per sentence for EFL learners and ESL learners who share the same language proficiency B1_2. Given the identical variables like topic, time limit and proficiency level, such differences are accountable to the language backgrounds of them. It is believed that those ESL learners are probably more inclined to use coordination phrases in their sentences while EFL learners tend to use them less, although those EFL and ESL learners are identified the identical proficiency level. Likewise, the obvious higher statistics of ESL learners over EFL learners in other syntactic measures can also be explained with their different preferences in writing which are not necessarily a result of proficiency difference.

4.2.3.2 Advantages of measures divided by sentences

Measures divided by sentences rather than clauses or T-units (see discussion in 3.2.6) seem to better signal proficiency levels. The previous data analysis suggests that whenever certain structures divided by clauses and structures divided by sentences are compared, the latter seems to be more indicative across the three groups whereas in some situations the former may fail to do so. For example, while be-copula structures per clause may fail to signal the difference between the three groups, be-copula structures per sentence is able to do so. Consequently, it is recommended that in future research measures divided by sentence can be used to replace the widely used measures divided by clauses.

4.2.3.3 Variation of more advanced participants

As noted in the data analysis, more advanced participants, including ESL learners and ENL writers, tend to show more variations in terms of

those syntactic complexity measures. For instance, the standard deviation for coordinate phrases per sentence for EFL learners is 0.18 while for ESL Learner and ENL writers the figures are 0.28 and 0.31 respectively. This is largely because more advance learners and writers are capable of using more varied structures or techniques in their writing to realize complexification while for most EFL learners they are more often than not bound by the perceived rules in writing.

4.2.3.4 Difficulty of applying T-unit-based measures in syntactic complexity research

As revealed earlier, T-unit-related measures, the long established set of measures for evaluating syntactic complexity is not quite satisfying in signalling syntactic complexity. Only half of the eight measures seem to be indicative of proficiency levels. Moreover, it seems difficult to generalize or categorize them compared with the ease of making judgement with the other category of measures. Such a complicated situation proves that those T-unit-related measures are not straight-forward and indicative of proficiency levels on the whole.

4.2.4 Pedagogical implications

Given the obvious link between proficiency and certain syntactic measures like sentence length and subordination-based measures as well as those measures divided by sentences, language teachers can adjust the teaching methods and revise the teaching material accordingly to help learners approximate the native writers. For instance, EFL students should be encouraged to use more complex nominals and more subordination/ coordination structures in order to produce longer sentences and realize

higher syntactic complexity, which generally will in return translate into high score in tests.

4.3 Correlation between syntactic complexity measures

Some syntactic complexity measures are found to be correlated with each other, indicating a possible causal relationship between them. This can be especially helpful for revealing how advanced ESL learners and ENL writers produce longer sentences or clauses. Some methodological implications and pedagogical implications can be drawn accordingly.

First, there is a strong correlation between subordination-based measures and global syntactic complexity measures among the three groups of participants. Number of dependent clauses per sentence and mean length of sentences show a correlation figure as high as 0.79 for the three groups as a whole. Naturally, we can infer that the increase of dependent clauses will increase the mean length of sentences or clauses per sentences considerably.

Second, coordinate phrases will also contribute to the mean length of sentences, as coordinate phrases per sentence show a quite high correlation figure with mean length of sentences. It merits attention that the correlation between coordinate phrases per clause and mean length of sentences is not so strong, partially because the increase or the drop of coordinate phrases per clause may not impact the sentence length directly.

Third, the use of complex nominals per sentence is also found to positively correlated to global syntactic complexity, including both mean length of sentences and clauses per sentence. It is reasonable to infer that the increase of complex nominals may positively influence the sentence length and number of clauses. Consequently, two global complexity measures

featuring sentence length and clauses per sentence are also affected. Moreover, according to the statistics of correlation, complex nominals may also contribute to the number of dependent clauses per sentence. Probably, some complex nominals may entail a dependent clause, which is consistent with the definition of dependent clause for this research.

Fourth, number of be-copula structures per sentence is positively related to the length of sentences and clauses per sentence as well as dependent clause. It is believed that be-copula structure has also contributed to mean length of sentences and number of dependent clauses which in turn results in increased ratio of clauses per sentence.

Last, mean length of clauses is positively related to coordinate phrases per sentence and complex nominals per clause. It is not difficult to infer that coordinate phrases and complex nominals can contribute to the length of clause since both of them are often included within clauses.

4.4 Topic effect on syntactic complexity

Topics in corpora were found to account for the differences between varietal types in some earlier studies (Danzak, 2011; Hundt & Vogel, 2011; Wulff & Römer, 2009). This may also suggest the possible effect of topic on syntactic complexity. The research findings provide support to this assumption since a strong topic effect on certain syntactic complexity measures is identified in this research.

On the whole, the topic on part-time job seems to help participants produce more complex sentences compared with the topic on smoking. This is especially true for the mean length of sentences, coordination-based complexity measures and phrasal complexity measures. The

subordination-based measures and it-cleft-related complexity measures, however, are not strongly influenced by topic effect.

As for sentence length, obviously, the three groups all produce longer mean length of sentences for the topic on part-time job. Further statistical analysis may suggest that in terms of sentence length, the more proficient the group is, the more vulnerable to be influenced by topic. In addition, coordination-based measures are also strongly influenced by topic. This is especially true for EFL and ESL learners since both groups exhibit significantly higher level of syntactic complexity when the topic is part-time job. This seems to suggest that learners, be they are EFL learners or ENL learners, are more inclined to be influenced by topic in their use of coordinate structures. In addition to the influence on sentence length and coordination structures, topic is also found to impact on the phrasal complexity, including all of the three phrasal complexity measures. Mean length of clauses and complex nominals in writings with the topic on part-time job is significantly higher than those with the topic on smoking.

The use of subordination-based measures and it-cleft-related complexity measures seems to be less sensitive to topic regardless of the topic in the three groups. This seems to indicate that the use of dependent clauses is relatively stable in the two topics. Partially due to the relatively smaller number of it-cleft structures, there is no observable difference of them across the two topics.

An important cause for the differences of syntactic complexity in the two topics is probably the attitude towards the argument. For the topic on part-time jobs, the vast majority of participants may have two contrasting

attitudes: support or refute. However, for the topic on smoking, almost all participants are against it in their writing. They almost unanimously criticize how harmful smoking can be while for part-time job people may evaluate both of its advantages and shortcomings. Besides, topic on part-time job may involve more personal experience, given the fact that all EFL and ESL learners are college students and half of the ENL writers are students. Probably people tend to produce sentences with higher syntactic complexity, for instance, longer sentences and more frequent use of coordinate phrases, when the topic is disputable and related to their personal experience. More specifically, when people have quite different opinions towards a topic and when they have experienced something related to the topic, they may be able to elaborate on the topic with more complicated language, which might result in more complicated syntactic structures. On the contrary, people tend to use less complicated language if the topic is not so disputable and familiar for them. In this regard, both disputableness of topic and familiarity with topic seem to contribute to the syntactic complexity, which can be tested in future research.

Based on the observation of topic effect of syntactic complexity, it is advisable for foreign language teachers to consider adopting certain topics to help learners produce more complex sentences. Preferably, those topics should be disputable in nature and should involve some personal experience of writers.

4.5 Chapter conclusion

This chapter offers further discussion on the result analysis in order to explain the result and draw implications. It is noted that certain syntactic

complexity features are indicative of proficiency. Based on those observations, methodological and pedagogical implications are proposed. Strong correlations between certain complexity measures are also identified and elaborated to account for them. In addition, topic does impact on certain complexity measures and the causes for it are also tentatively explained to offer pedagogical suggestions.

CHAPTER FIVE: CONCLUSION

5.1 Reflection on research findings

Despite the importance of syntactic complexity, there is a scarcity of corpus-based studies on it, much less studies on a comparison of syntactic complexity of EFL, ESL and ENL. This research has attempted to bridge this gap by conducting a detailed analysis of syntactic complexity in EFL, ESL and ENL groups. Following a multidimensional annotation scheme of syntactic complexity features, three comparable sub-corpora from the ICNALE have greatly facilitated the research process by providing reliable data. The study has to some extent demonstrated the great potential of corpus in studying syntactic features and the power the CIA in learner corpus research.

The original contribution of this study lies in its attempt to apply the corpus-based method to systematically examine the syntactic complexity of both EFL and ENL learners as well as ENL writers with the help of highly comparable datasets. During the examination, certain measures seem to be identified to be positive indicators of proficiency. Coupled with phrasal and coordination-based measures divided by sentences, global syntactic complexity measures and subordination-based complexity measures are found to be most indicative in identifying proficiency levels. Moreover, correlations between certain measures are also established tentatively in accordance with the statistical analysis. For instance, global complexity measures are found to positively correlate with subordination-based measures and the use of complex nominal structures while mean length of clauses is found to be positively associated with the use of complex

nominals and dependent clauses. Last, the topic effect on certain syntactic complexity measures is also explored, with topic on part-time job influencing mean length of sentences, coordination-based complexity measures and phrasal complexity measures as well as specific measures based on be-copula with adjective structures.

This study may shed light on the following aspects: Methodologically, this study may provide a useful example of examining syntactic complexity with annotated learner corpora and a certain set of complexity measures. Both automatic annotation and manual annotation are found to be useful in the data analysis. Pedagogically, the implications drawn from the research findings may help educators improve teaching methods and material accordingly, for instance, the influence of topic on syntactic complexity may help foreign language teachers choose more suitable topics to exert learners' syntactic complexity to their limit.

5.2 Limitations and future directions

Looking back, this research may also suffer from certain unavoidable shortcomings and may suggest some directions for future corpus-based studies at sentence level.

First of all, due to the nature of learner language, some ungrammatical sentences may be ambiguous and thus posing challenges to the annotation of those structures needed for this research. For instance, for the following sentence found in an EFL learner's writing, the identification of clauses may be problematic.

"In my perspective, college students have part time job is necessary."

Although such occasions are rare and generally limited to EFL data, it still deserves attention in this research. It is hoped that in future research the automatic annotation system can be further improved to better deal with learner data. Manual annotation is also necessary for identifying specific structures, although this requires more time and efforts. Automatic annotation and manual annotation can be combined to strike a balance between efficiency and accuracy.

Another notable limitation is that the writing samples in those datasets are relatively short writings with 200 to 300 words, which make some less infrequent syntactic structures less visible on the whole. Preferably, future learner corpora can consider including longer writing samples, say, 500 words or more for each sample while the number of participants should be ensured for the sake of representativeness.

As for the generalization of the research findings, it is also noted that in this research Chinese learners and Singapore learners are chosen as EFL group and ESL group respectively, which may result in the overgeneralization of the differences of the two learner groups. More varieties of EFL or ESL can be included in future research to improve the generalizability.

On a final note, to get a better understanding of how syntactic complexity develops among a certain group, it is sensible to collect some longitudinal data to capture the development process, which can further explain the developmental process of language progression. Such longitudinal research on language at syntactic level can be meaningful given the scarcity of such studies.

BIBLIOGRAPHY

Armstrong, K. M. (2010). Fluency, accuracy, and complexity in graded and ungraded writing. *Foreign Language Annals, 43*(4), 690-702.

Arthur, B. (1979). Short-term changes in EFL composition skills. In K. P. C. Yorio & J. Schachter (Ed.), *On TESOL' 79: The Learner in Focus* (pp. 330- 342). Washington, DC: TESOL.

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp & M. D ęz-Bedmar (Eds.), *Linking-up contrastive and learner corpus research* (pp. 35-53). Amsterdam: Rodopi.

Bardovi-Harlig, K. (1992). A second look at T-Unit analysis: reconsidering the sentence. *TESOL Quarterly, 26*(2), 390-395.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*(1), 17-34.

Becker, A. (2010). Distinguishing linguistic and discourse features in ESL students' written performance. *Modern Journal of Applied Linguistics, 2*, 406-424.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre?. *Reading and Writing, 22*(2), 185-200.

Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grades three to seven: syntactic complexity and genre differentiation. *Reading and Writing, 24*(2), 183-202.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1), 2-20.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly, 45*(1), 5-35

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

Borin, L., & Prutz, K. (2004). New wine in old skins? a corpus investigation of L1 syntactic transfer in learner language In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 67-88). Amsterdam: John Benjamins Pub.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks*. Princeton: Educational Testing Service.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education, 20*(1), 50-66.

Carlsen, C. (2012). Proficiency level--a fuzzy variable in computer learner corpora. *Applied Linguistics, 33*(2), 161-183.

Carter, R., & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge, England: Cambridge University Press.

Chen, M., & Zechner, K. (2011). *Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech.* Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics.

Connors, R. J. (2000). The erasure of the sentence. *College Composition and Communication, 52*(1), 96-128.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research*, 69, 176-183.

Crystal, D. (2008). *Dictionary of linguistics and phonetics*. Oxford and Malden, MA: Blackwell.

Danzak, R. L. (2011). The integration of lexical, syntactic, and discourse features in bilingual adolescents' writing: an exploratory approach. *Language, Speech, and Hearing Services in Schools, 42*(4), 491-505.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing* (pp. 155-165). Norwood, NJ: Ablex.

Davydova, J. (2012). Englishes in the outer and expanding circles: A comparative study. *World Englishes, 31*(3), 366-385.

de Haan, P., & van Esch, K. (2006). Assessing the development of foreign language writing skills: syntactic and lexical features. *Language and Computers, 60*(1), 185-202.

Deterding, D. (2010). *Dialects of English: Singapore English*. Edinburgh: Edinburgh University Press.

Dıaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. Paper presented at *Language forum,* 36(1 ·2). 139 ·154.

Dickinson, M., & Ragheb, M. (2009). Dependency annotation for learner corpora. Paper presented at the *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories* (TLT-8).

Dornyei, Z. (2005). *Psychology of the language learner: Individual differences in second language acquisition*. Mahwah: Lawrence Erlbaum Associates.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations. *International Review of Applied Linguistics, 47*(2), 157-177.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167-192). Amsterdam: John Benjamins.

Flowerdew, J. (2010). Use of signalling nouns across L1 and L2 writer corpora. *International Journal of Corpus Linguistics, 15*(1), 36-55.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299-324.

Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly,* 53-60.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course*: Routledge.

Gilquin, G. (2003). Automatic retrieval of syntactic structures: The quest for the holy grail. *International Journal of Corpus Linguistics, 7*(2), 183-214.

Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 55-78). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*(1), 97-114.

Granger, S. (1994). *From CA to CIA and back: an integrated contrastive approach to bilingual and learner computerised corpora.* Paper presented at the Languages in Contrast: Papers from a Symposium on Textbased Cross-linguistic Studies, Lund.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam & Philadelphia: Benjamins.

Granger, S. (2011). How to use foreign and second language learner corpora. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: a practical guide* (pp. 7-29). Chichester, West Sussex, UK: John Wiley and Sons.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *The International Corpus of Learner English V2* (Version 2. ed.). Louvain-la-Neuve, Belgium: Université catholique de Louvain.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL, 19*(03), 252-268.

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus- driven study of learner use. In M. Charles, D. Pecorari, S. Hunston & I. ebrary (Eds.), *Academic writing : At the interface of corpus and discourse* (pp. 193-214). London: Continuum.

Halliday, M. A. K. (1989*). Spoken and written language (2nd . ed.)*. Oxford: Oxford University Press.

Halliday, M. A. K., & Webster, J. H. M. A. K. (2004). *The language of science*. London: Continuum.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Hartmann, R. R. K., & Stork, F. C. (1972). *Dictionary of language and linguistics*. London: Applied Science.

Hasselgård, H., & Johansson, S. (2012). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. D. Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 33-62). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: theory and illustrations. *English Profile Journal, 1*(01), e5.

Hinkel, E. (2003). Simplicity without elegance: features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-301.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly, 18*(1), 87-107.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*: Cambridge University Press.

Hudson, R. (2009). Measuring maturity. In R. Beard, D. Myhill, M. Nystrand & J. Riley (Eds.), *The sage handbook of writing Development*.

Hundt, M., Denison, D., & Schneider, G. (2012). Relative complexity in scientific discourse. *English Language and Linguistics, 16*(02), 209-240.

Hundt, M., & Vogel, K. (2011). Overuse of the progressive in ESL and learner Englishes-fact or fiction?. In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 145-166). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3. Champaign, IL.: National Council of Teachers of English.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development, 35*(1), iii-67.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183-205.

Ishikawa, S. i. (2011). A new horizon in learner corpus studies: the aim of the ICNALE project. In S. I. G. Weir & K. Poonpon (Eds.), *Corpora and Language Technologies in Teaching, Learning and Research* (pp. 3-11). Glasgow, UK: University of Strathclyde Press.

Ishikawa, S. i. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (Vol. 1, pp. 91-118). Kobe, Japan: Kobe University Press.

Kern, R. G., & Schultz, J. (1992). The effects of composition instruction on intermediate level French students' writing performance: some preliminary findings. *The Modern Language Journal, 76*(1), 1-13.

Kirkpatrick, A. (2011). Learning English and other languages in multilingual settings: principles of multilingual performance and proficiency. *Australian Review of Applied Linguistics, 31*(3), 1-11.

Klein, D., & Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 3-10). Cambridge, MA: MIT Press.

Laporte, S. (2012). Mind the gap! bridge between world Englishes and learner Englishes in the making. *English Text Construction, 5*(2), 264-291.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-619.

Levy, R., & Andrew, G. (2006). *Tregex and Tsurgeon: Tools for querying and manipulating tree data structures.* Paper presented at the Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing, 11*(1), 5-21.

Little, D. (2007). The Common European Framework of Reference for Languages: perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645-655.

Lorenz, G. R. (1999). *Adjective intensification: Learners versus native speakers: a corpus study of argumentative writing*. Amsterdam: Rodopi.

Low, E. L. (2010). English in Singapore and Malaysia: similarities and differences. In A. Kirkpatrick (Ed.), *Routledge handbook for world Englishes* (pp. 229--246). London: Routledge.

Low, E. L., & Brown, A. (2005). *English in Singapore: An introduction*. Singapore: McGraw Hill.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME journal, 32*, 97-114.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua, 115*(11), 1619-1639.

Meurers, D., & Müller, S. (2009). Corpora and syntax. In A. Lüdeling & H. M. Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 44, pp. 920-933). Berlin: Walter de Gruyter Verlag, Kapitel.

Mukherjee, J., & Gries, S. (2009). Collostructional nativisation in new Englishes verb-construction associations in the International Corpus of English. *English World-Wide, 30*(1), 27-51.

Myhill, D. (2006). Designs on writing (2): designing sentences. *The Secondary English Magazine, 10*(3), 23-28.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research, 21*(4), 373-391.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. *Reading & Writing Quarterly, 23*(3), 287-309.

Nesselhauf, N. (2009). Co-selection phenomena across new Englishes parallels (and differences) to foreign learner varieties. *English World-Wide, 30*(1), 1-26.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578.

O'Donnell, M. (2013). UAM CorpusTool (Version 3.0). Retrieved from http://www.wagsoft.com/CorpusTool/download.html

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

Ortega, L. (2012). Interlanguage complexity: a construct in search of theoretical renewal. In B. Szmrecsanyi & B. Kortmann (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact Languages* (pp. 127-155). Berlin: Walter de Gruyter.

Osborne, J. (2011). Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics, 16*(2), 276-298.

Pennington, M. C. (2003). The impact of the computer in second language writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 287-310). Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., & Crystal, D. (1985). *A comprehensive grammar of the English Language*. Cambridge: Cambridge University Press.

Reid, J. M. (1993). *Teaching ESL writing*. Englewood Cliffs, NJ: Prentice Hall Regents.

Reilly, J., Zamora, A., & McGivern, R. F. (2005). Acquiring perspective in English: The development of stance. *Journal of Pragmatics, 37*(2), 185-208.

Reinhardt, J. (2010). Directives in office hour consultations: a corpus-informed investigation of learner and expert usage. *English for Specific Purposes, 29*(2), 94-107.

Rimmer, W. (2006). Measuring grammatical complexity: the Gordian knot. *Language testing, 23*(4), 497-519.

Rimmer, W. (2008). Putting grammatical complexity in context. *Literacy, 42*(1), 29-35.

Saville, N. (2010). The English profile programme: background, current issues and future prospects. *Language Teaching, 43*, 238-244.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Smart, J., & Crawford, W. (2009). *Complexity in lower-level L2 writing: Reconsidering the T-unit*. Paper presented at the Meeting of the American Association for Applied Linguistics, Denver, CO.

Song, M. (2006). *A correlational study of the holistic measure with the index measure of accuracy and complexity in international English-as-a-Second-Language (ESL) student writings*. (Unpublished doctoral dissertation). The University of Mississippi, Oxford, US.

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity. *JADT-04, 2*, 1032-1039.

Szmrecsanyi, B., & Kortmann, B. (2011). Typological profiling: learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (Eds.), *Exploring second-Language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 167-187). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420-430.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language, 31*(1), 101-122.

Tono, Y. (2009a). Corpus-based research and its implications for second language acquisition and English language teaching. In T. Kao & Y. Lin (Eds.), *A new look at language teaching and testing English as subject and vehicle* (pp. 155-173). Taipei, Taiwan: The Language Training and Testing Center.

Tono, Y. (2009b). Integrating learner corpus analysis into a probabilistic model of second language acquisition. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 185-203). London: Continuum Intl Publ Group.

Tono, Y. (2010). Learner corpus research: some recent trends. In G. Weir & S. Ishikawa (Eds.), *Corpus, ICT, and language education* (pp. 7-17). Glasgow: University of Strathclyde Publishing.

Vaezi, S., & Kafshgar, N. B. (2012). Learner characteristics and syntactic and lexical complexity of written products. *International Journal of Linguistics, 4*(3), pp. 671-687.

Van Rooy, B. (2011). A principled distinction between error and conventionalised innovation in African Englishes. In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: bridging a paradigm gap* (pp. 191-209). Amsterdam: Benjamins.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, *96*(4), 576-598.

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal*, *97*(S1), 1-20.

Weaver, C. (1996). *Teaching Grammar in Context*. Portsmouth, NH: Boynton/Cook Publishers.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing, 26*(3), 445-466.

Willis, D. (2003). *Rules, patterns and words: grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: measures of fluency, accuracy, & complexity*. Hawaii: University of Hawaii Press.

Wulff, S., & Römer, U. (2009). Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora, 4*(2), 115-133.

Xiao, R. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian Journal of English Language Teaching, 3*(2), 1-19.