STATISTICAL MODELLING OF LEXICAL AND SYNTACTIC COMPLEXITY OF ACADEMIC WRITING: A GENRE AND CORPUS-BASED STUDY OF EFL, ESL, AND ENGLISH L1 M.A. DISSERTATIONS

by

MARYAM NASSERI

A thesis submitted to
The University of Birmingham
For the degree of
DOCTOR OF PHILOSOPHY

Department of English Language and Linguistics
School of English, Drama and Creative Studies
University of Birmingham
October 2020

Abstract

This research is an interdisciplinary study that adopts the principles of corpus linguistics and the methods of quantitative linguistics and statistical modelling to analyse the rhetorical sections of MA dissertations written by EFL, ESL, and English L1 postgraduate students. A discipline-specific corpus was analysed for 22 lexical and 11 syntactic complexity measures using three natural language processing tools [LCA-AW, TAALED, Coh-Metrix] to find differences of academic texts by English L1 vs. L2 and to investigate the relationship between these linguistic indices. Structural factor analyses as well as the two statistical modelling methods of linear mixed-effects modelling and the supervised machine learning predictive classification modelling were then employed to verify the existing classification of the complexity indices, to explore their further dimensions, to investigate the effects of English language background and rhetorical sections on the production of lexically and syntactically complex texts, and finally to predict models that can best classify the group membership and the membership to the rhetorical sections based on the values of these measures. This investigation resulted in more than 20 specific findings with important implications for academic writing assessment of English L1 vs. L2, for academic writing research on rhetorical sections of English academic texts, for academic writing instruction especially materials development and syllabus designs in the EFL contexts, and academic immersion programmes, for the measure-testing and selection processes, and for methodological aspects of statistical modelling in corpus-based academic studies.

Key words: Lexical Complexity, Syntactic Complexity, Academic Writing, Statistical Modelling, Corpus Linguistics, NLP

To my husband,

Ryadh,

who dragged me out of my comfort zone,

and showed me

a whole new world!

Acknowledgements

I have always thought that writing the acknowledgements is the easiest and the hardest part of my thesis! The easiest because for once I can ignore all the rules of grammar, formality, punctuation, AND 'passive voice'. It's the hardest because human language can only convey gratitude in so many words as 'appreciate', 'thanks.' etc. So many people directly or indirectly contributed to this research; the ordering of their names here is only a limitation of language. My sincere gratitude goes:

To my parents Razieh and Alimorad: my mother who dragged me on the streets to school when I decided to quit school at the age of seven! And my father who taught me not to be afraid of anything (he went slightly overboard on this, I confess!). They persisted that I spend all my life learning and supported me along this path. They also collected all the EFL data for this study.

To my husband for his ongoing support, for giving a direction to my life, for (literally) forcing me to learn R and Python, and for caring for everything I do.

To my supervisors. Dr. Paul Thompson who patiently helped me throughout the PhD journey and supported me in my decisions and applications. He has been meticulously proofreading my drafts and showing me how to be precise and 'academic'. He gave me space to learn and grow as an independent researcher but has been always there for me when I needed advice, suggestion and feedback. Dr. Bodo Winter who was the main reason I adopted statistical modelling for my project. We engaged in very interesting and long discussions about statistics. He clarified some misunderstandings I've had about the fundamentals in statistics and gave great feedback on my results.

To the examiners of this thesis, Dr. Magali Paquot and Dr. Akira Murakami who will be spending many weeks reading this work, which will hopefully keep them isolated from COVID-19!

To Prof. Xiaofei Lu who was my main source of inspiration for studying lexical and syntactic complexity. He kindly and patiently answered all my questions and showed me how to modify the lexical analyser for my work.

To Prof. Jenifer Larson-Hall whose statistics book motivated me to pursue statistics and R further. She kindly answered my questions and encouraged me to continue this path.

To Prof. Philip McCarthy with whom I have had interesting discussions about the effectiveness of some measures of lexical diversity. He gave great insights about the usefulness of some measures and helped me to critically approach them rather than taking them for granted.

To Prof. Scott Jarvis for suggesting several measures of lexical diversity and how to critically assess them in my research.

To Coh-Metrix research team, especially Prof. Zhiqiang Cai (Carl) who gave me access to the full version of this analyser and shared important information about the computation method of a lexical diversity measure.

To Prof. Bethany Gray who helped me to better understand syntactic complexity and gave valuable suggestions about my research.

To the authors of several R packages: Dr. Max Kuhn, Prof. John Fox, Dr. Marvin Wright for clarifying some misunderstandings about the tests and for answering my R and statistics-related questions.

To other statisticians, the lavaan gmail group members, and SEMNET members who directed me towards better understanding the structural equation modelling and factor analyses, especially Prof. Edward Rigdon, Prof. Ryan Bowles, Dr. Sacha Epskamp, Cameron McIntosh, and Dr. Andy Liaw.

To Prof. Nader El-Bizri and Prof. Mark Smith who provided evidence for the use of rhetorical structures in early scientific works.

To all academic and non-academic members of staff at the University of Birmingham who helped me during the process of data collection and other administrative works.

To all the students who either shared their master’s dissertations or helped me in the laborious process of data collection.

To the anonymous reviewers of my journal articles and conference papers, for providing critical feedback on my manuscripts which greatly influenced this thesis.

And to the authors of all the books and articles that I have used in my study, for their contribution to my research.

Table of Contents

5. Research Design, Methodology, and Methods

6. Statistical Procedures, Results, and Discussions of the Findings

7. Concluding Remarks: Conclusions, Implications, and Suggestions for Future Research

List of Illustrations

List of Tables

List of Frequently-used Abbreviations

AILMRC: Refers to the rhetorical and organisational pattern of Abstract, Introduction, Literature review, Method & design, Results & discussion, Conclusion
ANC: American National Corpus
ANOVA: Analysis Of Variance
BAWE: British Academic Writing English
BNC: British National Corpus
CAF: Complexity, Accuracy, Fluency
CARS: Create A Research Space
CEFR: Common European Framework of Reference
CFA: Confirmatory Factor Analysis
CI: Confidence Interval
EAP: English for Academic Purposes
EFA: Exploratory Factor Analysis
EFL: English as a Foreign Language
ESL: English as a Second Language
ESP: English for Specific Purposes
IMRD: Refers to the rhetorical and organisational pattern of Introduction, Method, Results, Discussion
L1: First language
L2: Second language
L2SCA: L2 Syntactic Complexity Analyzer
LCA: Lexical Complexity Analyzer
LCA-AW: Lexical Complexity Analyzer for Academic Writing
LD: Lexical Density
LFP: Lexical Frequency Profiler
Log: Logarithm
LV: Lexical Variation
LS: Lexical Sophistication
MA: Master’s/ Master of Arts
MFUW: Most Frequently-Used Words
ML: Machine Learning
NLP: Natural Language Processing
NS: Native Speaker, in this thesis refers to English L1
NNS: Non-Native Speakers, in this thesis refers to non-English speakers
POS: Part of Speech
RA: Research Article
RF: Random Forest
SEM: Structural Equation Modelling
SLA: Second Language Acquisition
TAALED: Tool for the Automatic Analysis of Lexical Diversity
TEFL: Teaching English as a Foreign Language
TTR: Type-Token Ratio

1 Introduction

1.1. An Introduction to Linguistic Complexity

Theses and dissertations, as distinct genres of academic writing, are viewed as the "most significant piece of writing that any student will ever do" (Hyland, 2004, p. 134) and "the most sustained and complex piece of academic writing (in any language) they will undertake" (Swales, 2004, p. 99 on master's dissertations). Despite these observations about the importance of master's dissertations as the first serious and long scholarly pieces of scientific writing of students, the literature offers little insight about the complex linguistic processes and features that the writing of such texts entails. Comparative analyses of academic writing proficiency and performance of postgraduate students with different English language backgrounds is even more scarce. Little do we know, for example, about the predominant linguistic features of academic writing and dissertation sub-genres (e.g., rhetorical sections) written by English L1 vs. L2 students in different academic contexts (e.g., EFL vs. ESL). There is also little consensus on the reasons for proficiency disparities between these students regarding the production of high-quality texts expected to earn MA degrees in these contexts. Swales (1990, p. 188) for instance, observes that the analysis of research theses and dissertations has been "largely avoided, at least partly because of the daunting size of the typical text". With the advancement of computational linguistics tools and techniques, however, the analysis of long texts in large-scale corpora is no longer an impediment. Such tools and programmes that are fundamentally based on Natural Language Processing (NLP) techniques, as well as advanced statistical computing capabilities, provide objective and quantifiable measures of linguistic performance and proficiency. Statistical modelling of different genres of texts is a relatively new and exciting research area and a by-product of this advancement in text processing and statistical computation.

By adopting the methods of corpus linguistics and statistical modelling and by taking advantage of advanced NLP tools and methods, therefore, I set out to systematically investigate the differences of postgraduate students with different English language backgrounds regarding the production of lexically and syntactically complex texts in sub-genres or rhetorical sections of their master's dissertations to revisit the theories and studies of differences of English L1 vs. L2 texts. The use of advanced statistical modelling methods and predictive models also enables us to test the hypotheses regarding the assumed structures and the relationships between various linguistic constructs and measures, to examine the effects of English language background and rhetorical sections on the values of linguistic indices, and to detect strong lexical and syntactic predictors of rhetorical section and group memberships. Before delving into the significance of this research and the details of the objectives, it is essential at this point to define ‘complexity’ and ‘linguistic complexity’ and various approaches that help us understand and examine the linguistic complexity of master’s dissertations.

The term 'linguistic complexity' has found its way into various sub-disciplines and areas of research in linguistics. Linguistic complexity and its various components, constructs and measures of lexical, semantic, grammatical, syntactic, and morphological complexity, for instance, have been the focus of first and second language acquisition and development (e.g., Beers and Nagy, 2011; Norris & Ortega, 2009; Wolfe-Quintero, Inagaki, & Kim, 1998, etc), text readability, classification, and simplification (e.g., DuBay, 2004; Flesch, 1948; Vajjala, 2015), language impairment and decline, for example aphasia and agrammatism, dementia, language impairment in children, etc (e.g., Durán et al., 2004; Evert, Wankerl, & Nöth, 2017; Peristeri, Andreou, & Tsimpli, 2017), computational linguistics especially in the automation of linguistic complexity measures (e.g., Chen & Meurers, 2016; Lu, 2010, 2012; Kyle & Crossley, 2016, 2017, etc), text stylistics and stylometrics (e.g., Štajner & Mitkov, 2011) register variation studies (e.g., Biber & Gray, 2010, 2013, 2016), native language identification or the detection of author’s first language (e.g., Kyle, Crossley & Kim, 2015), second language acquisition research that investigates learners' proficiency levels, text comprehension and reading (Rayner & Duffy, 1986; Rawson, 2004), and research on writing quality (e.g., Beers & Nagy, 2009; Crowhurst, 1983, etc).

In the context of Second Language Acquisition (SLA), development, and proficiency research, linguistic complexity is frequently investigated with respect to two central constructs of ‘syntactic complexity’ and ‘lexical complexity’ (e.g., Kuiken, Vedder, & Gilabert, 2010; Szmrecsanyi & Kortmann, 2012; Wolfe-Quintero et al., 1998 among many others) which serve a wide variety of purposes, notably the analysis of L1 and L2 development and proficiency in learner corpora (e.g., in the works of Kim, 2014; Li, 2000; Lu, 2012; Shah et al., 2013; Vaezi & Kafshgar, 2012; Vidaković & Barker, 2009; Wolfe-Quintero, et al., 1998).

This diversity of research areas as well as the variation in research designs, objectives, and applications of linguistic complexity has led to different definitions and conceptual classifications of this term and consequently, different approaches to studies of linguistic complexity (e.g., Bulté & Housen, 2012; Dahl, 2004; Housen & Kuiken, 2009; Housen et al., 2019; Larsen-Freeman, 1997; Pallotti, 2015; Szmrecsanyi & Kortmann, 2012). However, two broad approaches to linguistic complexity can be detected in the various categorisations and definitions of this term in studies on Second Language Acquisition (SLA) and linguistic complexity. I address these broad areas as the 'systems view' and 'functional view' of linguistic complexity. In the following sections, I first give concise descriptions of what each approach to studying linguistic complexity entails and then I situate my research at the intersection of these areas and explain how various research questions and sections of this thesis contribute to our understanding of a more expansive picture of linguistic complexity in the context of SLA, corpus, and academic writing research.

1.2. Linguistic Complexity Based on Approaches to 'Complexity'

In order to demonstrate how the two broad mentioned areas of understanding linguistic complexity are indeed inter-related, it is a requisite to review how the term 'complexity' is defined, viewed, and addressed in various research areas.

The word 'complex' ('complexe' in French and 'complexus' in Latin) is composed of 'com' (together) and 'plectere' (to weave, to plait, to entwine) and is commonly defined as either 'a whole made up of multiple, different and connected parts', i.e., 'composite' (first recorded 17th century) or as 'complicated and difficult to understand and analyse' (first recorded 18th century), which is the opposite of 'easy' and 'simple' (Bastardas-Boada, 2017; Stevenson's 2010 dictionary entry of 'complexity', Oxford English Dictionary). Its derivative 'complexity' is consequently defined as composite nature and intricacy, the opposite of simplicity and simpleness (first recorded 18th century) and from 1794 used to mean "a complex condition" (Online Etymology Dictionary, 2019; Bastardas-Boada, 2017; Mufwene, Coupé, & Pellegrino, 2017; Stevenson, 2010, Oxford Dictionary of English). It generally characterises the state of anything comprised of multiple interconnected parts which interact in different or complicated ways. The term 'Complexity' also refers, to "diversity of forms, to emergence of coherent and orderly patterns out of randomness" as well as "to a significant flexibility" of switching among such patterns to attain the optimal ones in the designated context (Drożdż, Kwapień, & Orczyk, 2009, p. 1044).

However, there seems to be no singly unified definition for the term 'complexity' and different branches of knowledge appear to adopt different yet related approaches and definitions for it (e.g., computational complexity theory, Kolmogorov or algorithmic complexity, mathematical or Krohn-Rhodes complexity, etc). In scientific studies, for instance, Bar-Yam (2002) regards complexity and 'complexity science' as the study of “how parts of a system and their relationships give rise to the collective behaviors of the system, and how the system interrelates with its environment” (p. 2). The study of these interrelationships and behaviours is, therefore, the objective of 'complex systems theory'. Weaver (1948) among pioneers of complexity studies in science, classifies the concept into “organized complexity” and “disorganized complexity”. The former denotes a system with correlated relationships which behaves and interacts with other systems non-randomly and has emergent properties which could be approached via cross-discipline collaboration and be understood by computer modelling and simulation. Planetary orbits, for instance, could be considered as an organised system. The latter, on the other hand, is perceived as a system, such as behaviours of gas molecules in a space, where numerous elements (numerous-variable problems) interact in generally random ways and can be understood using probability analysis and statistical methods.

Three main definitions to ‘complexity’ are therefore central to most complexity studies. The first is related to the concept of ‘composite’ or a system made up of multiple inter-related parts (e.g., various internal inter-related yet distinct components, measures, and constructs) where different internal components in the system contribute to the overall complexity (see for example the discussions in Bar-Yam, 2002; Bastardas-Boada, 2017; Weaver, 1948). This first notion is reflected in Pallotti (2015) as "a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns" (p. 118). The second notion is related to the meaning of ‘complicated’ and ‘difficult to process, understand, and produce’, etc; this sense of complexity manifests itself in linguistic complexity studies e.g., in the context of cognitive complexity where the central theme is the discussion of processing load and the cognitive demands that a task imposes on the learner. This point will be further elaborated in section 1.2.3. The last meaning is related to the notion of ‘dependency on multiple external factors’ where complexity (e.g., the complexity of a given text) is a function of various/several other independent variables (e.g., the topic, genre, language background of students, age, gender, task types). For instance, a system is said to be complex when it depends on multiple factors. Regarding these external factors, Mufwene, Coupé, and Pellegrino (2017) argue that “complexity arises not just from how the different parts interact with each other but also from how they respond to external pressures of the environment, or the external ecology” (p. 3). In this latter sense, variation in any of these external variables affects the complexity level of a text for instance. These meanings of the term ‘complexity’ will be elaborated in the following two sections.

1.2.1. Systems View of Linguistic Complexity

The systems view of linguistic complexity has its roots in the theories of ‘complex systems’ and ‘complexity science’ as discussed above. The science of complex systems introduced a plethora of avenues to problem-solving and explanation of complex and nonlinear systems, as well as new theories and approaches such as dynamical systems theory, chaos theory, and complexity theory which explain the behaviours of components in complex systems. Such systems are called 'complex' not only because they are composed of numerous and various components, but also because of the constant interactions (actions and reactions) of these components in a sometimes unpredictable way. That can partly explain why the behaviour of the entire system cannot be understood by solely analysing the behaviours of its individual components. The study of complex systems also gave rise to an array of new and innovative ideas and approaches to problems in different fields of study. Perhaps one of the most influential of them is the ‘chaos complexity theory’ which found its way into interdisciplinary language studies such as Second Language Acquisition (SLA).

The much-celebrated Chaos Complexity Theory, for example, features characteristics such as ‘complex’, ‘nonlinear’, ‘dynamic’, ‘unpredictable and chaotic’, ‘open’, ‘sensitive to initial conditions’, ‘feedback sensitive’, ‘adaptive and self-organising’, and ‘seeking strange attractors’ as are explained below.

Larsen Freeman (1997) exquisitely drew a comparison between such a system and language (particularly the interlanguage system), claiming that language is essentially a complex system; its complexity is attributed to its numerous sub-systems such as syntax, semantics, morphology, etc as well as their interdependency and interactions in a fashion that the whole language cannot be comprehended just by examining its individual parts in isolation. Likewise, this complex system is nonlinear in that “the effect is disproportionate to the cause” (p. 143), such as a small rolling pebble causing an enormous avalanche. In this view, the process of language learning is nonlinear as well in that components of a language are not learnt linearly, one at a time. The dynamism of language can be best appreciated when viewing it not as a set of fixed standard rules and products, but as an active process of growth and change (such as the developing nature of L2 learner's internal grammar) and its diachronic nature. Complexity theory also proposes that by using the language, we are changing it every time.

The interlanguage system like other complex systems is further affected by and sensitive to initial conditions which could set the future behaviour of the system. Minute initial changes may lead to dramatic changes. Such systems are feedback sensitive as well. They self-organise themselves based on the positive or negative feedback they receive and hence adapt and modify themselves towards order, complexity and maturity. The case in point is similar to when L2 learners absorb the positive feedback to modify their interlanguage grammar to that of the target language grammar. Language/interlanguage as a complex system also gravitates towards and settles into attractor states – the unmarked states. These fields of attraction permits the language to accommodate infinite new inputs to its finite phonological and morphosyntactic rules. Different L2 learners, on the other hand, might be tamed by their L1s strange attractors and settle, for instance, for different pronunciations of the same L2 word.

Two other approaches to the studies on complexity science exist, according to Bastardas-Boada (2017): one stream focuses on computation and modelling of complex systems and the other on epistemological and philosophical studies. Housen et al. (2019) have also elaborated on other approaches to complexity science with a systems view in second language research, such as Dynamic Systems. Among the SLA studies, several works adopted the dynamic systems or the usage-based perspective, such as Verspoor, Schmid, and Xu (2012) as well as Vyatkina's (2012) developmental studies of second language writing mainly at the inter- and intra-individual variabilities.

Therefore, there are various characteristics of complex systems (e.g., the discussion on the interlanguage system above), and various types of studies relevant to complex systems. However, when it comes to its application in linguistics, the mentioned approaches are mainly 'conceptual' rather than a 'math-oriented' systematic method used in traditional complexity science. The salient point in the discussion of linguistic complexity with a systems view is the emphasis on the number and types of components in a system, the structure of the linguistic constructs (as dependent variables), the interaction between these components (e.g., the relationship among the constructs and their representative measures), and the emphasis on how variation in one component/construct (and its constituent measures) affect other constructs and measures. In this sense, linguistic complexity and its sub-domains of lexical, syntactic, semantic, grammatical, morphological, and pragmatic complexity constructs and their constituent measures have inter-related yet distinct properties in a way that, for example, larger values of lexical complexity of a given discourse may correlate with higher semantic complexity and possibly correlate with larger syntactic complexity values. I will further discuss this point regarding the objectives of this thesis in section 1.2.3. In chapter six I will demonstrate how the computational and statistical modelling approach to systems view of linguistic complexity furthers our understanding of the structure and behaviours of various lexical and syntactic constructs.

1.2.2. Functional View of Linguistic Complexity

In what I call the ‘functional view’ of linguistic complexity, complexity is not defined in isolation, but is taken as a function of task (e.g., task type and condition), genre (e.g., academic writing), rhetorical features (e.g., rhetorical sections of articles, book, and dissertations), English language background (e.g., EFL, ESL, English L1), topic (e.g., disciplines and sub-disciplines and research areas) and sample size (e.g., the length of texts in terms of tokens) among other variables. Many studies on linguistic complexity in the context of SLA, corpus and writing research, for instance, aim to contextualise linguistic complexity and its constructs like lexical, syntactic, and morphological complexity and investigate how the variation in such contexts and variables affect the complexity level of a given discourse. The complexity of a text is said to be a function of the genre when, for example, it is supported that the lexical density of a research article abstract is higher than that of a descriptive essay. It can also be argued that, for instance, syntactic complexity is a function of the task in that certain types of tasks elicit certain types of structures or the complexity level of a discourse (e.g., Michel et al.’s 2019 study on the “effects of task type on morphosyntactic complexity” of L2 writing). Lu (2011) also found that syntactic complexity levels of L2 writers can be affected by time (e.g., more allocated time for writing).

There is a wealth of research that has adopted this approach to understanding linguistic complexity (with various terms/nomenclature to address this approach), especially regarding the first and second language acquisition, development, and proficiency (e.g., Beers & Nagy, 2009, 2011; Biber & Gray, 2010, 2016; Ellis, 2003; Housen & Kuiken, 2009; Housen et al., 2019; Lu & Ai, 2015; Olinghouse & Wilson, 2013; and Skehan, 1992, 2009b among many others). Various fine-grained complexity terminologies (e.g., task complexity, L2 complexity, cognitive complexity, etc) are used by researchers to further examine this broad functional approach and the effects of tasks, cognitive processing, etc on the values of objectively-defined linguistic complexity measures.

When task is taken into account, linguistic complexity as Ellis (2003) defines it, is regarded as “the extent to which the language produced in performing a task is elaborate and varied” (p.340). This elaboration and variety aspect of linguistic complexity, Housen and Kuiken (2009) state, is gauged based on the language features such as patterns and structures as well as syntactic, lexical, and morphological features. Housen and Kuiken (2009) however, view linguistic complexity based on two distinct areas. One is the complexity of the interlanguage system that I discussed under the ‘systems view’; the other is the inherent complexity of linguistic features and forms which they call ‘structural complexity’.

In the context of second language acquisition and development, the two notions of ‘task complexity’ and ‘task difficulty’ are frequently addressed (e.g., in Housen & Kuiken, 2009). Task difficulty, as Skehan (1992, 1996) proposes, is affected by language factors (lexical and syntactic complexity and range) and cognitive factors (e.g., familiarity with materials, task type, and the amount of mental processing required). Along similar lines, Robinson (2001, 2005) proposes that ‘task complexity’ relates to the structure of the task and the cognitive complexity of tasks, i.e., the cognitive demands that the task places on the learner, and ‘task difficulty’ which is perceived by the learners, and is affected by the learners’ ability (e.g., proficiency and intelligence) and affective factors (e.g., anxiety and motivation).

With respect to 'cognitive complexity', Barker and Pederson (2009) consider the ‘processing load’, whereby the complex statements are those which are difficult to process in terms of ‘informational content’ as well as production and comprehension of the utterances. Although speakers are practically restricted by certain grammatical choices, they have control over degrees of simple and complex constructions. They distinguish between meaningfully complex constructions versus complex forms: a single word can be complex in meaning (e.g., having several unrelated meanings) while a longer utterance could be complex in form (e.g., based on the length of the utterance, etc). We should as well acknowledge, as they state, that complexity is relative: a two-word statement is simpler in structure than a sentence, yet more complex than a single-word utterance. Bulté and Housen (2012) equate cognitive complexity with relative complexity or the difficulty in ‘processing or internalizing’ linguistic features; i.e., a feature is called complex when it is ‘cognitively taxing’ for language users.

Regarding these variables, two lines of research have can be noticed. The first line is when ‘cognitive factors’ are taken as independent variables or when the effects of tasks’ cognitive complexity on the production of linguistically complex discourse are investigated. The second line is when linguistic features are taken as independent variables and the effects of their internal/structural complexity are examined (e.g., whether long vs. short words and structures can affect learnability and cognitive processing).

Other researchers have investigated the effects of other variables, such as genre, registers, topics, and language backgrounds on the complexity level of a given discourse or corpus. Biber and Gray (2010, 2013, 2016), for instance, examined the effect of different genres and register variation on the syntactic and grammatical complexity of texts, using various grammatical and syntactic structures and indices. The function of genre and the effect of different genres on the variation in lexical and syntactic complexity values are also closely studied in other works with various research designs (e.g., Beers & Nagy, 2009, 2011; Olinghouse and Wilson, 2013; Stewart and Grobe, 1979). The effect of language background on the complexity level of texts written by English L1 vs. L2 students is also examined in Lu and Ai (2015) as well as Crossley and McNamara (2010) among other similar works. Likewise, the effect of topic on the production of syntactically complex academic texts was investigated in Yang, Lu, and Weigle (2015). Yoon (2017) also investigated the effects of topic and proficiency level on the values of lexical, syntactic, and morphological complexity indices obtained from college students' argumentative essays. Detailed accounts of these and similar other studies will be provided in chapter three. A review of the prominent works in these areas suggests that these works have acknowledged that the complexity level of a given discourse or corpus is a function of other variables; therefore, complexity measures are taken as dependent variables to probe the effects of genre and register variations, task (e.g., task type, task complexity, condition, etc), first language background and target language, as well as age, gender, proficiency level, and disciplinary variations on the values of various linguistic complexity indices.

1.2.3. Linguistic Complexity at the Crossroads of Systems View and Functional View

The degree to which the definitions of 'complexity' and the approaches to measuring 'linguistic complexity' are inter-related depends on the research design and questions.

As discussed in section 1.2.1, the central theme in the systems view of linguistic complexity is the inter-relationships between multiple components of a system, the structure of these components, and whether the variation in one component correlates with the variation in other components in the same or opposite direction. Taking a text or a corpus as a system, this study opts for quantifying lexical and syntactic complexity of master's dissertations produced by postgraduate students. The performance and proficiency (differences) of three groups of students with different English language backgrounds will be determined regarding the production of lexically and syntactically complex academic texts (e.g., as the complexity level expected to earn master's degrees) in respective academic contexts. Within the systems approach, a text is said to be lexically complex when, for instance, various lexical constructs such as lexical density, diversity/variation, and sophistication and their representative measures receive large quantifiable values, when these constructs are verified to be distinct and the constituent measures of each receive higher positive correlation with each other than with the measures of another construct, and when the values (e.g., numerical objective values) of one construct correlate positively with the values of another construct (e.g., when a text with higher lexical diversity also receives larger lexical sophistication values). These investigations will be carried out in chapter six using correlation tests and structural equation models that, as will be explained in detail, consist of confirmatory and exploratory factor analyses. The implications of adopting this view and the related findings will be further discussed in the final chapter, section 7.3.

The second approach that I address as the ‘functional view’, is assumed in the literature in investigations of the effects and functions of various factors/variables (e.g., the role of task, task conditions, task planning, genre, rhetorical functions, academic context, L1s, etc) and concepts and theories (e.g., cognitive complexity and the effects of processing load and difficulty, informational content, etc) in the production of complex spoken and written discourse. This line of studies are manifest in the works of Alexopoulou et al. (2017), Ellis (2009), Housen and Kuiken (2009), Johnson, M. D. (2017), Kormos (2011), and Sadeghi and Mosalli (2012) among others.

In this study by adopting the functional approach to linguistic complexity, lexical and syntactic complexity will be contextualised to gauge the effects of task (dissertation writing), sub-genres and rhetorical sections of dissertations, and English language backgrounds of students (EFL, ESL, and English L1) on the production of lexically and syntactically complex texts, while controlling other independent variables of age, gender, and sub-disciplinary variations. A diverse set of statistical modelling methods will be employed in chapter six to find the best models that can explain the highest amounts of variation regarding the effects of rhetorical sections and language backgrounds on complex texts, as well as the best models which can pinpoint strong lexical and syntactic predictors of rhetorical section and group memberships.

As mentioned, there are several different classifications and/or understanding of different approaches to linguistic complexity that serve different types of studies. The salient point in this discussion is that there are overlapping areas among these broad approaches and the ways they could be implemented in research design in linguistics studies, and therefore these areas are not mutually exclusive. In the final chapter I will revisit this point to show how the two approaches that I adopt in this study could be seen as complementary.

1.3. Linguistic Complexity and Proficiency

Most linguistic complexity indices/measures have been proposed as developmental indices for SLA studies that gauge second language proficiency and development in an objective way (Larsen-Freeman, 1978, 1983; Wolfe-Quintero et al. 1998). The underlying assumption in the field, as Wolfe-Quintero et al. (1998) states, is that “second language learners write more grammatically and lexically complex sentences as they become more proficient” (p. 4). This assumption is also reflected in Ortega (2003) and Housen et al.’s (2019) discussions of complexity in the SLA literature. Acknowledging the presence of context-specific and individual variability, the researchers in the past two decades set off to examine this assumption. Although the research designs of these studies differ, generally strong positive correlations are reported between most of these linguistic complexity measures and higher levels of linguistic proficiency based on the programme levels, proficiency test scores, experts’ subjective rankings, and specific proficiency levels like the CEFR, Common European Framework of Reference (see for instance Doró, 2008; Kim, 2014; Lu, 2012; Ortega, 2003; Treffers-Daller, Parslow, & Williams, 2016; Yang, Lu, and Weigle, 2015 among others) or comparisons with English L1 students (e.g., Ai & Lu, 2013; Gonzalez, 2013; Linnarud, 1986; Lu & Ai, 2015; Pietilä, 2015). The cumulative findings suggest that such complexity indices could be used as indicators of proficiency as will be discussed in detail in chapter three. This point will be further discussed in 3.3. in the discussion of criterion validity and the relationship between certain lexical and syntactic complexity measures and proficiency.

A point that deserves clarification is the answer to the question ‘are more complex texts necessarily better?’. The short answer would be ‘no’ when admitting the myriad of factors that affect the quality of texts, for instance, the organisational aspects, semantics, cohesion, and adherence to the norms of discourse communities, just to name a few. However, a multifaceted long answer seems indispensable.

In the context of first and second language acquisition and writing research, as well as linguistic proficiency, performance, and development, the use of various linguistic complexity indices (e.g., lexical, syntactic and/or grammatical, morphological complexity measures) is meaningful only in comparative studies where multiple groups of learners/students with various language backgrounds and linguistic proficiency levels are compared (e.g., with each other or with a higher proficiency group like an English L1 group or expert writers) or in developmental studies where the development of certain linguistic features is traced for one or more groups of learners. In this context, many studies have reported strong positive correlations between larger values of one or more of these linguistic complexity indices and the holistic ratings of writing quality and/or other indicators and predictors of writing quality, proficiency and development (e.g., Lu & Ai, 2015; Mancilla et al., 2015; Yang et al., 2015; Wolfe-Quintero et al., 1998 among many others). This is particularly the case in advanced and academic writing that, as Hinkel (2003) emphasises, the use of simple lexical and syntactic structures is viewed as a 'severe handicap'. Many researchers have underlined the use of more complex linguistic structures (e.g., compared to English L1 students and academics or high-quality texts as specified by holistic rating) and a more compressed style of academic writing not only an indication of overall writing quality and linguistic proficiency, but also as meeting the expectations of discourse communities (see for example the discussions in Biber &Gray, 2010; Crossley & McNamara, 2012; Hinkel, 2003; Laufer & Nation, 1995; Lu, 2012, 2017).

Another aspect of this argument is that the required complexity level of discourse is highly dependent on its rhetorical function, as well as on genre expectations and conformity to discipline-specific norms. Some examples are the simplified, explanatory, less complex, and tutorial-like writing style in textbooks vs. the condensed, and packed-with-terminology writing style in journal articles. The required complexity level of a text also depends on its audience/readers, its objectives, and the communicative aspects of that text. For instance, considering the specialised audience of theses/dissertations (and other specialised academic texts like journal articles) that are already familiar with basic concepts, terminology, and discipline-specific structures, it seems more efficient, as Biber and Gray (2010) argue, to be able to extract the most amount of information in relatively short segments of (longer) texts in a relatively short time. These types of texts (i.e., mainly advanced academic texts) would inevitably become more dense, embedded, and sophisticated in nature, and hence of higher complexity levels.

I end these discussions here with brief notes on the distinction between complexity and comprehensibility and/or readability. Since there is no upper or lower bound for the values of various linguistic complexity measures, it is more plausible to think of linguistic complexity (specifically lexical and syntactic complexity) as a continuum stretching from very low complexity (e.g., very simple discourse like a child's story) to very high complexity (e.g., very sophisticated, philosophical and/or abstract discourse) both in terms of the difficulty of production and perception of lexical and syntactic structures (although this definition is not the main focus of this research) and in terms of the number of constituents (e.g., how many linguistic elements/structures are used/understood in a discourse), their type and variety (e.g., various lexical and syntactic constructs) and their degrees of interconnection (e.g., if higher values of a linguistic variable or construct correspond to higher values of another variable or construct). In this view, complexity and readability/comprehensibility may seem to be opposing forces. However, the crucial point to consider is that each discourse type on this spectrum and its corresponding complexity and readability values serve different purposes and functions based on the genre and field-specific expectations, proficiency levels and the audience of the discourse as mentioned earlier. Academic writing, for instance, tends to fall towards the more complex part of the spectrum where high linguistic complexity values seem to denote higher linguistic proficiency and development (see, for example, Ai & Lu, 2013; Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998), partly because addressing complex academic concepts and ideas requires longer and more complex syntactic structures and lexical units.

The cumulative findings in the scholarly body of work regarding the relationship between certain linguistic complexity indices and proficiency will also be considered as a main yardstick to interpret the results of this study regarding the quantitative values of the selected measures. A detailed discussion of lexical and syntactic complexification as well as the effectiveness of these complexity measures will be presented throughout chapter three.

Several arguments and premises have been so far presented (and will be further elaborated in the following chapters). These include the context of this study's research (i.e., academic writing of postgraduates), the comparative nature of the study (i.e., comparing the ESL and EFL groups to the English L1), the role of rhetorical functions, genre and disciplinary expectations, specialised audience, as well as the consensus of most-cited researchers on the positive relationship between the higher values of these linguistic complexity indices and the writing quality, and linguistic proficiency and performance (except the discussions on syntactic coordination that will be discussed in chapter three). On the grounds of these arguments and premises, more lexically and syntactically complex texts in this study are considered to be of higher quality written by more proficient writers.

1.4. Research Gaps and the Significance and Objectives of this Study

There are four main research gaps and areas that motivate the main objectives of this study. These research areas will be briefly discussed in the following sections and more detailed discussions will be provided throughout the literature review sections as well as the final chapter in the implications of the findings.

1.4.1. Lexical and Syntactic Complexity Differences in English L1 vs L2 Specialised Academic Writing Corpora

Considering the English language as a first or a second language reflects a host of research studies which stress differences and similarities of L1 and L2 writing with regard to aspects of grammar, syntax, and lexis. Compared to the multitude of SLA studies that examine English L1 vs L2 text differences regarding general English language corpora both outside and in the context of academia, sparse and infrequent studies have systematically investigated linguistic complexity differences of specialised academic writing corpora, particularly regarding different genres and sub-genres.

Among general SLA studies, Silva (1993) points out the distinct nature of L2 writing and argues that L2 writers' sentences include "fewer but longer clauses, more coordination, less subordination, less noun modification, and less passivization" (p. 668). He comes to this general conclusion that L2 writers' texts "exhibited less lexical control, variety, and sophistication" (p. 668). Similarly, the key findings of Crossley and McNamara (2009) indicate that L1 writers use more abstract and hierarchically connected words, more polysemous words, word hypernymy, meaningful words, and more causal verbs than L2 writers. Several other studies acknowledge differences among general English academic writers: L2 academic writers "employ excessively simple syntactic and lexical constructions" than native English speakers (Hinkel, 2003, p. 275), show a limited supply of lexicon which lead to vague and less complex texts compared to native English writers (Carlson, 1988 cited in Silva, 1993; Leki, 1991; Read, 2000), and dramatically different usages of tenses and voice relative to L1 writers of English (Hinkel, 2004). These general English studies in the context of academia have been also carried out to investigate the effect of L1s (e.g., Crossley & McNamara, 2012; Lu & Ai, 2015) mainly in argumentative and narrative writings on such proficiency differences.

These and numerous other studies have emphasised the role of large-scale corpus-based studies for the identification of prominent linguistic patterns for comparative analyses of L1 vs. L2 texts in naturally occurring corpora as a starting point for English L2 instruction. Hinkel (2003), for instance, claims that in academic texts, syntactic and lexical simplicity is a 'severe handicap', and that certain syntactic and lexical features should be 'explicitly targeted in instruction' to English L2 writers. Both Pienemann (1985) and Pica (1985) also agree that the analysis of simple vs. complex linguistic features in L2 production contributes to curriculum and syllabus development, in that focus is directed to facets of language, such as 'syntactic regularities' and expanded 'lexical repertoire', that can be explicitly taught.

Descriptions of the linguistic profile of students' academic writing in terms of the frequency of use and distinct patterns of certain lexical and syntactic structures, as well as comparisons across the groups, are reported to contribute to L2 writing research and the measure-selection processes for further research, as well as for materials development, syllabus design, the students' self-study and awareness-raising in graduate and postgraduate programmes in the EFL and ESL academic settings (e.g., the discussions in Hinkel, 2003, 2004; Lu, 2012; Lu & Ai, 2015; Pica, 1984, 1985; Shahriari, Ansarifar, & Pishghadam, 2017; Silva, 1993).

Despite the use of different numbers and types of linguistic complexity measures, corpora, sample sizes, etc, these cumulative findings indicate that undergraduate texts, mainly argumentative writings of English L1 and L2 students differ regarding the diversification of lexis, the amount of sophisticated words, subordination, and phrasal complexity features. However, little do we know whether such differences persist at more advanced levels such as postgraduate writings, especially in discipline-specific and genre-specific texts. Among the few such studies, only Pietilä's (2015) investigated the conclusion sections of MA dissertations using a few lexical complexity measures and found significant differences between English L1 and L2 groups regarding the use of sophisticated words. Swales (2004, p. 99) describes theses and dissertations as "the most sustained and complex piece of academic writing (in any language) they will undertake". MA dissertations are usually the first serious scientific writing for most postgraduate students, especially in EFL academic contexts. In an attempt to revisit these theories and studies on L1 vs. L2 proficiency and performance differences of English texts at the postgraduate level, the present study, therefore, examines a discipline-specific specialised academic writing corpus of MA dissertations regarding a large set of lexical and syntactic complexity indices.

1.4.2. EFL vs ESL Academic Writing: Lexical and Syntactic Complexity in Specialised Corpora

In the context of general SLA studies, the term ESL (English as a Second Language) has either been used almost interchangeably with the term EFL (English as a Foreign Language), e.g., in Lu (2012) and studies reviewed in Silva (1993), or denoted English L2 students/learners with various L1s that were born and raised in an English-speaking country/community predominantly with English L2 families or the adult ESL learners who immigrated to an English-speaking country (e.g., in Joye, 2004). Following the tradition of Bley-Vroman (1990, p. 5), the term EFL in this context is specified as 'learning' English language in a non-English-speaking context (i.e., "the conscious learning of explicit rules", and ESL as ‘acquisition’ of English in an English-speaking context (i.e., “unconscious internalisation of knowledge” in an English-L1-mainly society). Even though the distinction between ‘learning’ and ‘acquisition’ in the context of academia requires systematic pedagogical investigations, examining textual differences in the academic writings of the two groups could be facilitated by large-scale corpus-based studies such as this study.

Comparing the academic writing of EFL and ESL students – i.e., those who pursue further education at a university in an English-speaking country – is a major underinvestigated area particularly regarding specialised and genre-specific academic texts. This distinction is mainly based on the academic context where the texts are produced. In an EFL setting English is not the primary language of education and communication (e.g., natural and authentic use of English) and therefore attending English classes is one of the main ways of exposure to English as a foreign language (see Nayar, 1997 for example). Academic writing in such settings is rarely taught outside academia. Derakhshan and Karimian Shirejini (2020) investigated the challenges of academic writing in Iranian EFL academic settings and emphasised that the limited use of English outside academia, the test-centred teaching and learning practices, and unfamiliarity with genre expectations and rhetorical structures are major issues that affect writing proficiency of students.

ESL students, however, can be regarded as former EFL students who transitioned to the ESL settings mainly as part of postgraduate academic immersion programmes sharing the same materials, syllabi, lecturers, and resources with their English L1 peers. In the case of the present study, the ESL students have moved to the UK solely for the purpose of postgraduate studies (either MA courses or both MA and PhD programmes) mainly as part of short- or long-term ESL immersion programmes developed/designed in their home countries. Since many of these students have received funding/studentships from an English L2 setting/university/institute, they are expected to leave the UK after graduation and/or receiving their postgraduate degrees, and therefore returning to an EFL setting. It is essential, therefore, to find out whether ESL students who benefit from academic immersion programmes and shared academic curriculum, materials, and modules with English L1s produce more linguistically complex texts, especially lexically and syntactically complex texts, than their EFL peers.

In her research synthesis of L2 writing proficiency, Ortega (2003) treated EFL and ESL academic contexts as separate variables and argued that the scholarly body of research in this area shows that “L2 competence may proceed more slowly and might develop less fully in foreign language than in second language instructional settings” (p. 498). She concluds that syntactic complexity differs systematically in the L2 writings of EFL and ESL students and postulates that these two contexts comprise “distinct L2 populations” (p. 512).

Masgutova and Kormos (2015) for instance investigated the role of academic ESL programmes in the UK and found that ESL students’ lexical sophistication and variety significantly improved. A similar developmental study of ESL students is conducted by Bulté and Housen (2014) where these students produced more syntactically complex texts by the end of an intensive academic English programme. It might not, however, be possible to derive a definite conclusion about the superiority of either setting as many other factors such as the role of students’ first language or socio-cognitive factors are involved in the proficiency of EFL and ESL students. Nevertheless, consistent and strong patterns emerging from a large enough dataset could be a reasonable indicator of proficiency differences in terms of the production of complex lexical and syntactic structures.

In this project, I will investigate dissertations written by Iranian EFL students. The academic writing and/or dissertation writing courses in the EFL settings (e.g., Applied Linguistics and TEFL disciplines) in universities in Iran mainly revolve around writing processes and mechanics of writing (e.g., Esmaeili & Esmaeili, 2015) and organisational aspects of writing (Sadeghi & Shirzad Khajepasha, 2015) – often following the simple traditional pattern of organisation described in Thompson (1999) – and less attention is paid to raising awareness and explicit teaching of linguistic structures which contribute to the total linguistic complexity of the texts compared to the texts produced by English L1 writers (e.g., dissertations/theses). EFL graduates are often left to themselves to consult already-published dissertations of EFL and English L1s to grasp the spirit of academic writing and linguistic structures required for writing a dissertation as the most important and lengthy scholarly work of graduate students.

Scanning dozens of dissertations, I noticed how this negligence leads to overly simple lexical and syntactic structures in EFL dissertations and/or filling up the required structural gaps by frequent and extended direct quotations from the works of other scholars. This could in certain cases hinder the communication of ideas since complex ideas could be better put across via complex syntactic structures and lexical units (see for instance the discussions in Hinkel, 2003). Assessment issues arising from employing simple constructions and vocabulary have already been discussed in Hamp-Lyons (1991) and Davidson (1991). Other scholars have also raised their concerns about the lack of native-like or high-quality texts in linguistics-related disciplines in the EFL academic settings such as Iran and its impact on their academic achievement (Karimnia, 2013; Maleki & Zangani, 2007; Sahragard, Baharloo, &

Soozandehfar, 2011; Sadeghi & Shirzad Khajepasha, 2015). Consequently, many researchers (e.g., Karimnia, 2013; Sadeghi & Shirzad Khajepasha, 2015) have called for English L1 academic writing data to be incorporated into such academic writing and thesis writing corpus-based studies for comparative analysis and understanding of the nature of English L1 vs. L2 production of linguistic features.

Despite decades of research and the building of language corpus resources in Europe, corpus linguistics is a relatively new field of study in Iran. However, recent years have witnessed a number of corpus studies emerging from researchers in applied linguistics to explain various linguistic phenomena in corpora of the naturally occurring language use as well as describing learners' linguistic performance. Particular attention is given to academic textbooks and academic writing genres such as research articles (Farvardin, Afghari, & Koosha, 2012; Gholami, Mosalli, & Bidel Nikou, 2012; Jalali & Ghayoomi, 2010; Jalilifar, Firuzman, & Roshani, 2011; Khany & Khosravian, 2013; Moiinvaziri, 2012; Salmani Nodoushan & Khakbaz, 2011; and Zarei Chamani et al., 2012) which reflects the increasing demand for academic writing studies. Among these corpus-based studies, Gholami, Mosalli, and Bidel Nikou (2012) for example examined lexical complexity in abstracts of research articles in hard and soft sciences.

Compared to the host of studies on research articles as products of expert academic writing, little work has been done on master's theses/dissertations, the exception being those which describe the language of dissertation sections (Abdollahzadeh, 2011; Jalilifar & Dabbi 2012; Jalilifar & Vahid Dastjerdi, 2010). Such studies are recommended to clarify, among other merits, whether the present linguistic knowledge of Iranian master's students is of a proficiency level comparable to that required from English L1 academic writers. The incorporation of the findings of such studies into the syllabus design and materials development processes of MA programmes, therefore, could have a significant impact on bridging the proficiency gaps early on and before the students find themselves struggling to produce high-quality texts (e.g., research articles in journals with high impact factors) as doctoral, post-doctoral and academic researchers (see for instance the discussions in Flowerdew, 2007 on the struggles of English L2 researchers to publish scholarly materials, and Flowerdew, 2015 on ERPP or English for Research Publication Purposes).

As discussed, research on the academic writing differences of EFL and ESL students regarding lexically and syntactically complex texts is an underinvestigated area. To my knowledge, there is no study so far that has examined such differences in postgraduate and specialised academic writing corpora. In this study, the performances of these two groups will be compared to revisit the hypotheses in the literature concerning the possible role of ESL academic immersion programmes in attaining native-like linguistic proficiency. The specific answers to the related research questions concerning the possible effects of the ESL and EAP academic programmes will be provided in chapter six, section 6.8.2; this point will be further discussed in the conclusion chapter, section 7.3 together with relevant findings in previous research and with suggestions for materials development processes in the EFL academic contexts as well as the inclusion of short and long term EAP and ESL academic programmes especially thesis/dissertation writing.

1.4.3. Research on Linguistic Features of Sub-genres or Rhetorical Sections of Specialised Academic Writing Corpora

This research was also prompted by two trends in academic writing studies, namely register variation studies and the research line on genre and rhetorical expectations of academic writing, especially regarding the linguistic features which have gained momentum with the works of notable scholars as will be discussed. Despite numerous works on genre moves and rhetorical structures of various academic writing genres and sub-genres, little has been done to describe academic writing as an overarching genre of writing and its various sub-genres and rhetorical sections regarding various lexical, syntactic, morphological, and grammatical features and constructs. The research in these areas can be even more informative when academic writing is contextualised based on the task, topic, discipline, and regarding the effects of rhetorical sections, writers’ L1s and English backgrounds, age, gender, etc (e.g., Biber & Gray, 2013; Flowerdew, 2017; Hinkel, 2004; Hyland & Shaw, 2016; Lu & Ai, 2015; Thompson, 2002 among other similar works). Lu, Casal, and Liu (2020) likewise emphasise the importance of form-function relationships and the linguistic realisations of various rhetorical functions in academic writing instruction at advanced stages.

The findings of this line of research could also be incorporated into automatic text classification and identification systems/models (see e.g., Shehan et al. 2010) for identifying sub-genres and rhetorical sections of very large corpora. Shehan et al. (2010) for instance incorporated a multiplicity of lexical, grammatical, and syntactic measures in text classification models and systems to distinguish various types of texts based on these linguistic features. The process of text classification involves assigning tags to various categories, e.g., linguistic features of interest, using natural language processing methods. It has important applications in sentiment analysis, spam detection, web search, information retrieval, ranking, document classification, and text complexity level analysis as discussed earlier. As one of the central themes of the present thesis, I will generate two models of lexical and syntactic complexity of dissertation texts written by English L1 vs. L2 postgraduates using a text classification method. I will further discuss at length how feature classification models as specified by rhetorical sections of dissertations (as already-labelled texts) can identify and predict prevalent lexical and syntactic features of rhetorical sections in postgraduate academic writing (chapter six).

In chapter four, I will describe various linguistic and genre features of the main rhetorical sections of advanced and specialised academic texts such as research articles and dissertations/theses. In light of previous research, I will then discuss the necessity of expanding the IMRD rhetorical structure (Introduction, Method, Results, Discussion) to include abstract and conclusion sections as de facto rhetorical sections of most theses/dissertations based on communicative goals, organisational patterns, and genre moves. In this work, for the first time, I investigate whether various rhetorical sections serving as sub-genres of postgraduate academic writing (MA dissertations) can be (distinctly) characterised by certain type and amount of lexical and syntactic features as used by students with different English language backgrounds. Using a set of global syntactic complexity measures and a large set of lexical complexity measures, I examine the extent of variability among the rhetorical sections regarding the type and distribution of these structures and to find out how much of this variability can be attributed to a main text-intrinsic characteristic (i.e., various rhetorical sections distinguished based on rhetorical and genre features and expectations) and to one main text-extrinsic factor (i.e., the groups of students based on English language backgrounds and academic contexts).

Statistical modelling methods are particularly powerful and valuable yet barely-used methods for undertaking studies like this research (see the discussions in Gries, 2015, 2019; Levshina, 2015; Winter, 2019). The detailed discussions in these works as well as the findings of this study are testimonies to the capabilities of these methods in obtaining predictive models of prominent lexical and syntactic features of rhetorical sections of MA dissertations written by students from three different English language backgrounds. It is against this backdrop of recent advances in statistical modelling and NLP methods (e.g., the ability to automate procedures for large-scale text analysis, handling collinear variables, scalability, ranking top predictors as text features, etc) that this research takes place.

1.4.4. Measure-testing Process of Lexical and Syntactic Complexity Indices Using Robust Statistical Modelling Methods

Traditionally, linguistic proficiency has been investigated through qualitatively analysis of texts such as learner essays by human raters. With the increasing number of English learners and the availability of large numbers of texts/corpora, however, this is no longer a viable option for assessing large learner corpora. Acknowleding the limitations, Doró and Pietilä (2015) emphases the importance of automated essays scoring systems for such assessments; they list studies that show high correlations between the results of these systems and human raters. A multitude of measures/indices have been, therefore, proposed to objectively quantify linguistic complexity and its over-arching constructs and dimensions of lexical and syntactic complexity using natural language processing (NLP) tools as alternatives to the qualitative analysis of texts by human raters. Despite the ease with which writing researchers can now parse and examine texts, several additional challenges are also acknowledged. A noticeable one is the presence of a multitude of linguistic measures proposed by quantitative linguists and writing assessment researchers to gauge these linguistic constructs. The mainstream NLP tools analyse hundreds of different measures of linguistic complexity. Acknowledging the role of other linguistic and non-linguistic variables, I limit the scope of this research to specifically focus on lexical and syntactic constructs and measures that are frequently reported in the literature to contribute to our understanding of linguistic proficiency of advanced and academic English texts. In chapter five, section 5.3, I will discuss the measure-selection process for this study in detail.

Two main issues arise from the multitude of proposed measures in the literature. As will be discussed in chapter five, among the set of lexical and syntactic complexity measures investigated in this study, many have been validity- and reliability-tested; however, few measures have not been thoroughly investigated. This necessitates the process of criterion validation, i.e., testing these less-studied measures against the well-established and reliable measures as indicators and predictors of proficiency. A second issue is the presence of several similarly-calculated measures to quantify a construct. For instance, among 22 indices of lexical diversity as a construct of lexical complexity that are investigated in this study, several indices are computed based on similar methods, e.g., based on logarithm or based on word-strings/segments. These many measures are formulated as a remedy to the text-length dependency of the ratio of types (unique non-repetitious words) to tokens (all words) as will be discussed in detail in the following chapter. Some of these proposed measures are different adaptations of one quantification method for overcoming this problem but to my knowledge, previous works have not investigated a large set of such closely-related indices to examine the effectiveness of each index in pairs/groups of related indices in capturing lexical and syntactic complexity differences, especially in genre-specific and field-specific academic corpora. These two issues will be investigated in this study in the measure-testing process consisting of several statistical tests and three robust statistical modelling methods of structural equation modelling, linear mixed-effects modelling, and random forest supervised machine learning as the predictive classification modelling. This is to obtain a more expansive picture of lexical and syntactic features of postgraduate academic writing in the MA dissertations written by three groups of students with different L1s and different English language backgrounds and to obtain a small set of distinct measures to quantify each construct of lexical and syntactic complexity. In chapter two, I will further describe these constructs and the measures that are proposed in the scholarly texts to quantify them.

1.4.5. Research Questions and Objectives

In light of the mentioned research gaps and objectives, the following 13 research questions as classified into four groups are specified. After detailed discussions of the results in chapter six, the answer to these research questions will be provided in section 6.8.

Group A of research questions (answered in 6.8.1) deals with the measure-testing process and examines the effectiveness of the 22 lexical and 11 syntactic measures in capturing differences of academic texts investigated in this study, the relationship between these measures and their overall and specific structures, and the best indicators and predictors of linguistic proficiency differences and text classification. The five specific questions are:

> **A1.** How do the selected lexical and syntactic complexity measures compare with and relate to each other as indices of quality of academic texts at the postgraduate level in the whole corpus of this study? Is the construct-distinctiveness of these lexical and syntactic categories (section 6.4) confirmed with this corpus of MA dissertations (section 5.2)?
>
> **A2.** To what extent do the selected lexical and syntactic complexity indices in this study fall into the current categories of lexical and syntactic constructs proposed in the literature (section 5.3.1)? What new structures are detected regarding this study's corpus of academic texts?

A3. Which lexical and syntactic constructs and measures can better capture differences in academic texts produced by three groups of postgraduate students (see details in 5.2.1 and 5.3.1) and what are the overall lexical and syntactic indicators of linguistic proficiency and performance as specified by the differences of group with different English language backgrounds (section 6.3)?

A4. What are the overall lexical and syntactic predictors of linguistic proficiency and performance of the groups (section 6.7)?

A5. Which of the lexical and syntactic indices explain the largest amount of variation in each dataset in the whole corpus as explained by mixed-effect models (section 6.6)?

Group B of research questions (answered in 6.8.2) deals with the comparisons of academic writings of the three groups of EFL, ESL, and English L1 postgraduate students. Three questions are, therefore, formulated as:

B1. Which group of students produced the most linguistically-complex texts, e.g., more lexically and syntactically complex texts (i.e., with larger values of each and/or all of the lexical and syntactic complexity measures and constructs selected in 5.3.1)?

B2. To what extent do the EFL and English L1 students/groups differ regarding the production of lexically and syntactically complex texts overall and specifically (e.g., based on the six rhetorical sections)? Do any such differences have implications for EFL academic writing practices?

B3. To what extent do the ESL students who benefit studying in the UK academic setting perform better than their EFL peers who study English in a non-English-speaking context, and to what extent do the ESL students' performances approximate the English L1 group considering the effect of the shared academic setting (i.e., academic programmes, materials, syllabi, and immersion in an English-speaking academic context)? Do any such differences have implications for ESL academic immersion programmes?

Group C of research questions (answered in 6.8.3) deals with the prominent linguistic features (e.g., patterns in lexical and syntactic constructs) in six rhetorical sections of master's dissertations. The two related questions ask:

> **C1.** What are the overall (dominant) lexical features of each of the six rhetorical sections of MA dissertations in terms of the lexical constructs of density, diversity and sophistication of the whole corpus? What are the top lexical predictors of each of the six rhetorical sections produced by all three groups combined?
>
> **C2.** What are the overall (dominant) syntactic features of each of the six rhetorical sections of MA dissertations in terms of the syntactic constructs of the length of production units, amount of subordination, amount of coordination, and degree of phrasal sophistication in the whole corpus? What are the top syntactic predictors of each of the six rhetorical sections produced by all three groups combined?

Finally, **group D of research questions** (answered in 6.8.4) examines two types of predictive statistical modelling based on regression and classification, as will be explained in detail in 6.6 and 6.7 respectively.

> **D1.** What are the effects of groups (English language background as English L1, EFL, and ESL) and rhetorical sections (the six sub-sections of MA dissertations), and their additive and interaction effects on the values of 22 lexical and 11 syntactic complexity indices? What are the best-fitting models which can explain the largest amounts of variations for these measures?
>
> **D2.** How accurately can we classify the groups of students based on the values of 22 lexical and 11 syntactic indices obtained from the analysis of academic texts (all six rhetorical sections combined)? What are the specifications of the best predictive models of group membership?
>
> **D3.** How accurately can we classify each of the six rhetorical sections of MA dissertations in this study's corpus based on the values of 22 lexical and 11 syntactic indices of the three groups of postgraduate students? What are the specifications of the best predictive models of membership to rhetorical sections?

1.5. An Overview of the Structure of this Thesis

The discussions in the next chapter will be both descriptive and critical in nature to present the most frequently-used indices of lexical and syntactic complexity along with the multitude of definitions, measurement criteria and quantification methods that have been proposed and used in the SLA, corpus, and academic writing literature. In chapter three I report on the findings of the main studies which have analysed one measure or a set of lexical and syntactic complexity measures to situate my research among closely-related research studies. Chapter four is dedicated to a principled survey of the structure and specifications of the main rhetorical sections of academic writing, especially theses/dissertations and research articles. Chapter five begins with the main theoretical premises and methodological issues behind this research to set the scene for the research questions and hypotheses which initially prompted this project. I will then set out to describe the details of the data collection and corpus construction processes along with comprehensive discussions on the measure-selection processes. The full account of statistical procedures and interpretation methods along with critical discussions of the findings compared to the results of previous studies will be presented in chapter six. This is followed by the concluding remarks of the findings in chapter seven with implications and applications of the findings, and a description of limitations and delimitations to set the direction for future studies with the same/similar research design.

2 Lexical and Syntactic Complexity Constructs: Their Representative Measures, Definitions, and Quantification Methods

2.1. Overview

Sections 2.2. and 2.3 and their sub-sections are dedicated to detailed descriptions of lexical and syntactic complexity respectively. Each section begins with an overview of the field, the definitions of major terms and constructs and the significance and applications of lexical and syntactic complexity in various types of research and studies. Each sub-section then concentrates on one main construct and its quantifiable measures and structures, and critically addresses these measures considering the evaluations and recommendations of experts. Lexical and syntactic complexity are multi-dimensional aspects of linguistic performance and proficiency. Throughout this thesis, the term 'construct' will be used to refer to various dimensions of lexical and syntactic complexity, i.e., the constructs of lexical density, diversity, and sophistication and the syntactic constructs of subordination, coordination, and phrasal complexity. In chapter five, I categorise lexical diversity to several sub-constructs based on specific quantification methods. The terms 'measure' and 'index' will be used interchangeably to refer to various ways these constructs could be quantitatively calculated.

The present chapter along with the studies that will be reviewed in the next chapter set the scene for the investigation of these constructs and measures based on the systems view and functional view that were discussed in the previous chapter. This is to examine (in chapter six) the relationship among these linguistic complexity constructs and measures and the ways that the amount of variation in each construct and measure affect other constructs and measures (systems view) and to contextualised these measures based on genre, task, English background of writers, etc to see how the number and type of these measures affect text-intrinsic and text-extrinsic factors (functional view).

2.2. Lexical Complexity Constructs and Measures: Terms, Definitions/Specifications, and Quantification Methods

Lexical complexity is a multidimensional aspect of lexical proficiency that is frequently used in the literature on first and second language acquisition and development, writing research (both learner and academic writing), and corpus-based studies of lexical performance and proficiency (differences) in learner or specialised corpora (e.g., Housen, Bulté, Pierrard, & Van Daele, 2008; Lu, 2012; Wolfe-Quintero et al., 1998). These studies often analyse one or more of the constructs of Lexical Density (LD), Lexical Sophistication (LS), and Lexical Diversity or Variation (LV) and their constituent measures to examine the quality of writing, the impact of task complexity and certain pedagogical interventions on the production of lexical features, to build lexical profiles of learners in terms of productive vocabulary knowledge and use, and to obtain quantifiable values to assess proficiency levels.

Lexical complexity and lexical richness are the two main nomenclatures that serve as the umbrella terms to encompass a range of quantifiable lexical indices in the mentioned studies. For the most part, these two terms seem to be used interchangeably, i.e., as an umbrella term that covers the mentioned lexical constructs mainly in cross-sectional studies of linguistic performance and proficiency (e.g., Gonzalez, 2013; Kim, 2014; Li, 2000; Lu, 2012; Shah, Gill, Mahmood, & Bilal, 2013; Vaezi & Kafshgar, 2012; Vidaković & Barker 2009). Read (2000) for instance, uses the term ‘lexical richness’ as an umbrella term for four constructs or components of lexical variation, sophistication, density, and the number of errors, addressing the effective vocabulary use in good writing. Likewise, Malvern et al., (2004, chapter 9) refer to the term lexical richness as encompassing several inter-related constructs of lexical diversity, sophistication (.e., the use of rare words), length of the words and texts, and the absence of vocabulary errors. For Ménard (1983), however, lexical richness is simply the number of word types in fixed-sized texts, comprising of the indices such as monosemic rate which is the use of monosemic (word types with only one meaning) rather than polysemic words as markers of concise writing. A number of other works also used the term ‘lexical richness’ as an umbrella term to analyse lexical density, diversity/variation, and sophistication (Laufer & Nation, 1995; Lu, 2012; Šišková, 2012; Kim, 2014) lexical density and variation (Linnarud, 1975), as well as lexical diversity/variation and sophistication (Vermeer, 2000; Daller & Phelan, 2007).

In some studies on the other hand, ‘lexical complexity’ seems to be adopted as a specific component in the CAF studies that examine the Complexity, Accuracy, and Fluency of a discourse or a corpus as indices of L1 and L2 proficiency and development (e.g., Chandler, 2003; Housen & Kuiken, 2009; Skehan, 1989; Wolfe-Quintero, Inagaki, & Kim, 1998; Zareva, Schwanenflugel, & Nikolova, 2005). The ‘complexity’ in this line of work based on the CAF triad consists of syntactic, grammatical, and morphological complexity. As mentioned earlier, ‘lexical complexity’ is either used as an umbrella term to analyse lexical density, diversity, and sophistication (Lexical Complexity Analyzer: Lu, 2012); lexical diversity and sophistication (Kol & Schcolnik, 2008); Lexical diversity and density (Vaezi & Kafshgar, 2012); and lexical diversity (Thomas, 2005; Johnson, Mercado, & Acevedo, 2012; Mazgutova & Kormos, 2015).

Despite the variations in the use and specification of lexical complexity and richness, in this study, I continue to use them interchangeably (in chapter three in the review of related studies and chapter six for interpreting the results) as umbrella terms subsuming several constructs, notably the three constructs of lexical density, diversity, and sophistication that are investigated in this thesis. In the introduction chapter, I have already elaborated on the term ‘complexity’ and the rationale behind using the term ‘lexical complexity’ in this regard. In the final chapter, I will also discuss how the findings of this study regarding these constructs and measures relate to the three approaches to linguistic complexity studies.

The three main constructs of lexical density, diversity and sophistication are frequently reported as reliable indicators and/or predictors of linguistic proficiency, performance, and development (e.g., Crossley & McNamara, 2010, 2013; Engber, 1995; Jarvis, 2013; Kim, 2014; Lu, 2012; Mazgutova & Kormos, 2015). Over the last few decades, various indices have been proposed to measure lexical knowledge and use and to describe the lexical profile of a corpus. There are some other proposed *ad hoc*, research-specific, and less-frequently-used lexical indices of ‘lexical originality’ or ‘lexical individuality’ as will be discussed again in 2.2.3, as well as ‘lexical specificity’ (Biber, 1988) that is considered as a characteristic of academic texts with precise words and high lexical variation; this latter index is calculated as the mean word length together with the TTR (type-token ratio) in a text’s first 400 words.

Nonetheless, some specific measures that will be discussed in this chapter have withstood scholarly criticisms and remained as staple factors affecting lexical production in the context of first and second language acquisition and writing research. In the sections that follow, I give detailed specifications of the three main constructs that are to be investigated in this study and their constituent measures along with the definitions and quantification methods proposed and used in various theoretical and research studies on first and second language acquisition, corpus-based, and writing research. The review of related literature suggests a variety of research designs and objectives for the use of these indices; in chapter five, however, I will discuss in detail the reasons for selecting the set of relevant measures (e.g., the indices that are more suitable for examining lexical performance and proficiency differences) for investigation in this study throughout the measure-selection section in 5.3.1.

2.2.1. Lexical Density

The term ‘lexical density’ is believed to have been introduced by Ure (1971) and is described as the proportion of lexical items to the total number of items or tokens (mainly a token-token ratio) in a written or spoken discourse (Halliday, 1985; Laufer & Nation, 1995; Johansson, 2008). These lexical items (also called ‘content words’ or ‘open class words’) comprise nouns, verbs, adjectives, and most adverbs; they are distinguished from the grammatical words (also called ‘function words’ or ‘closed classes words’ such as particles, prepositions, pronouns, determiners, and conjunctions) which serve to convey grammatical relationships and whose meanings depend on their functions in a sentence/unit of production. Simple as it appears, distinguishing between lexical and grammatical words proved to be challenging early on (Arnaud, 1984; Linnarud, 1975; Mendelsohn, 1981). This is mainly due to the issues with the boundaries between content and function words and multi-unit words. For example, a two-word item such as ‘turn up’ could be considered one lexical item (Halliday, 1985) or a lexical and a grammatical one (Ure, 1971). The scholars also have conflicting views regarding the lexical words and the role of derivatives, lemmas, and proper nouns just to name a few. To add to this complexity, the criteria for considering a word as a ‘lexical word’ are also not agreed upon in the scholarly works. Halliday (1985) and Lu (2012), for instance, have a stricter criterion for including verbs; they excluded modal verbs and the auxiliary verbs of ‘be’ and ‘have’. Likewise, for considering ‘lexical adverbs’, O’Loughlin (1995) included adverbs of time, manner, and place as ‘lexical’ items; Lu (2012) and Engber (1995) considered adverbs with the -ly suffix as well as the adverbs with an identical form of adjective in this category. The lack of a consensus has led to a practice of defining the criteria locally and when reporting the results. One advantage of lexical density, however, as Malvern et al., (2004) argue, over the type-token based measures (as will be discussed in the next section) is that most lexical density indices are token-token ratios and hence not affected by sample size.

Various forms of lexical density, though very infrequently used, are reported. Vajjala (2015) for instance, used several types of lexical density measures based on part-of-speech (POS) tags as their proportion to the total number of words or tokens, e.g., noun density (as the proportion of nouns and proper nouns to all tokens), adjective density (the proportion of all adjectives to tokens), pronoun density, verb density (non-modal verbs/all words) as well as the proportion of verb types to the total number of sentences, such as modal verb density (the proportion of modal verbs to the total number of sentences), VBN-tag density (verb be, past participle verbs/ total sentences), VBD-tag density (verb be, past tense/total sentences), VBG-tag density (verb be, gerund or present participle/ total sentences), and VBP-tag density (verb be, singular, present, non-3rd/ total sentences), among others. These latter indices based on the POS are usually used as indices in readability assessment (for the details see the discussions in Vajjala, 2015). Adjective density in Bates et al., (1988), however, is the proportion of adjectives to the content words, which in Lu (2012) is considered as adjective variation/diversity.

While some scholars view lexical density as an aspect of a text’s lexical diversity/variation(e.g., Stamatatos, Fakotakis, & Kokkinakis, 2000; Štajner & Mitkov, 2012), other researchers such as Lu (2012) have confirmed the construct-distinctiveness of these two in a large-scale corpus of oral narratives. This construct-distinctiveness will also be empirically examined in this study (section 6.2.3). In this chapter, section 2.2.5, I will further elaborate on these distinct constructs from theoretical and conceptual points of view and in light of previous studies.

2.2.2. Lexical Diversity/Variation

A seemingly related concept to lexical density is ‘lexical diversity’ or ‘variation’ which is often defined as the variety or range of different words in a text (Johansson, 2008; Housen, et al., 2008; Malvern et al., 2004), or to put it precisely, “phonologically-orthographical different word forms” that are representative of the size of vocabulary knowledge (Housen et al., 2008, P. 3). Lexical diversity sometimes has appeared in studies with other names as well: ‘lexical range and balance’ (Crystal, 1982) and ‘verbal creativity’ (the use of TTR in Fradis, Mihailescu, & Jipescu, 1992). Some researchers such as deBoer (2014) regard lexical diversity as a distinguisher between active and passive vocabulary to determine proficiency in first and second language acquisition. Others (Noyau & Paprocka, 2000; Dewaele & Pavlenko, 2003) link lexical diversity to productivity in descriptive or communicative tasks: it accounts for the amount of detailed lexical items an advanced learner may use to describe an event as opposed to general words used by beginners. In section 2.2.5, however, I argue that the use of general vs. advanced words does not necessarily correspond to the use of varied/diverse words (as commonly understood by lexical diversity).

Measuring lexical diversity, however, has been even more challenging. Since the idea behind this construct is to assess the variation / the use of non-repetitious vocabulary in a text, the most frequently-used and criticised method for calculating it is the type-token ratio (TTR).

This basic measure calculates the proportion of types (unique words) to the total number of tokens (all words) in a text. Two major problems immediately arise. One is whether to include the morphological variants of a lemma (inflections, derivations, and [formation of] compounds) as separate types. This decision could disturb the balance of types and tokens in favour of types, and consequently affect the results of lexical diversity. This problem holds true for any other measure that is based on the TTR as well. It can also be argued that the knowledge of inflected forms and derivatives of a word does not necessarily mean the knowledge of a diverse range of vocabulary in terms of non-repetitiousness.

The next difficulty with this calculation method is that TTR is highly text-length dependent. When measuring long texts, McCarthy and Jarvis (2007) explain, the beginning words are more likely to be new (i.e., types) and subsequent words are more likely to be the repetition of those types. Therefore, in the face of overwhelming number of tokens, the type-token ratio does not accurately reflect the pace of the new types as the text progresses (i.e., the ratio becomes too small for very long texts). This is partly dependent on the type of text. The next reason for this is that as the text length increases, the number of tokens increases due to the repetitive nature of function/grammatical words, but the number of types (unique words not used up to that point) does not increase with the same ratio. This problem has been extensively discussed in the literature (e.g., deBoer, 2014; McCarthy and Jarvis, 2010; Durán et al., 2004 among many other works). TTR is therefore, not suitable for comparing texts of varying lengths. Consequently, several mathematical and computational alternatives have been offered. Table 2.1 demonstrates 37 lexical diversity indices in the literature used in various research designs. These alternatives include logarithm-based measures, indices based on word-strings/segments, and measures based on the TTR of word classes. This table only provides a quick overview of these measures. The indices that will be investigated in this study will be described in more detail in chapter 5, throughout the measure-selection process in section 5.3.1. To see the extended description of the rest of these measures, refer to the citations in this table.

Despite the predominance of measures based on type-token ratios, several twentieth-century studies have used alternative indices of proportion, such as type-type ratio (as will be discussed more in the next section) and type-utterance ratio (TUR), for instance in Yoder et al.'s (1994) research on the prorated number of lexically free words (word types that are used in at least two varied combinations of words) and Richards's (1990) study on the use of the ratio of auxiliary words to hundred structured utterances.

Table 2.1. Some alternatives to Type-Token Ratio (TTR) and other indices used to measure lexical diversity/variation

Lexical Diversity Indices	**Specifications and Quantification Methods**
Linnarud’s LV	Linnarud (1975): $LV = (V_{lex} \times 100) / N_{lex}$
Mendelsohn’s LV	Mendelsohn (1981): $LV = V_{lex} \times 100 / N$
Arnaud’s LV	Arnaud (1984): $LV = V_{lex} / N$
Corrected TTR or CTTR	Carroll (1964): $T / \sqrt{2N}$
Index of Guiraud, also called RTTR or Root TTR	Guiraud (1954): $T/ \sqrt{N}$
Advanced Guiraud	Daller et al. (2003): Advanced (or rare or sophisticated) Type / √Token also used as a sophistication index
Yule’s K index/constant	Yule (1944): A measure of repetition based on the probability of a type in a random selection of two noun tokens: $K=10^4\times [(\Sigma^{x}_{x=1} fx\, X^2) - N] / N^2$
CR or Contiguity Rating	Perkins (1994): A token-token ratio measure of repetitiveness in language disroders
LRD or Limiting Relative Diversity	Malvern et al. (2004, p.148): It compares the diversity of different word classes; squre root of division of diversity of one word class to another
Brunet’s W index	Tweedie and Baayen (1998): $W= N^{V-a}$
Michéa’s M index	Michéa (1971): ratio of hapax dislegomena (v_2) to the total number of types in a text: $M= V/v_2$
Sichel’s S index	Sichel (1986): The notational inverse of Michéa’s M index: $S = v_2/ V$
Orlov’s Z index	A log-based measure based on Zipf’s law; it depends on the frequency of the most common word (in Malvern et al., 2004, p. 36)
Rubet’s K index	Dugast (1978): LogV/(LogLogN)
Somer’s S index	Somers (1966): (LogLogV)/(LogLogN)
Herdan’s C index, also called Bilogarithmic TTR or LogTTR	Herdan (1960): log V / log N

Lexical Diversity Indices	Specifications and Quantification Methods
Uber’s U Index	Dugast (1978): (log N) ^{2} /(log N– log V)
The Maas a^2 index	Maas (1972, cited in McCarthy & Jarvis, 2010) : Notational inverse of Uber index: Log N – log V (N) / Log ^{2} N
D Measure	Proposed by Malvern and Richards (1997), calculated in Voc-D programme (McKee et al., 2000)
HD-D or Hypergeometirc D	McCarthy and Jarvis (2007): Based on the hypergeometric distribution function
Measure of Textual Lexical Diversity (MTLD)	McCarthy (2005): The mean length of word strings that maintain a predetermined TTR
MSTTR or Mean Segmental TTR	Johnson (1944): Averages the TTR from all fixed-size segments of the texts/word strings
MATTR or Moving-Average TTR	Covington and McFall (2010): TTR for fixed-length successive moving winows (word strings) of a text
Lexical Word Variation	Linnarud (1986): T_{lex} / N_{lex}
VV1 or Verb Variation type I	Harley and King (1989): T_{verb} / N_{verb}
SVV1 or Squared VV1	Chaudron and Parker (1990): T^{2}_{verb} / N_{verb}
CVV1 or Corrected VV1	Wolfe-Quintero et al., (1998) as an adaptation of Carroll’s CTTR method: T_{verb} / √2 N_{verb}
VV2 or Verb Variation type II	McClure (1991): T_{verb} / N_{lex}
NV or Noun Variation	McClure (1991): T_{noun} / N_{lex}
ADJV or Adjective Variation	McClure (1991): T_{adj} / N_{lex}
ADVV or Adverb Variation	McClure (1991): T_{adv} / N_{lex}
MODV or Modifier Variation	McClure (1991): (T_{adj} + T_{adv}) / N_{lex}
NDW or Number of Different Words	Miller (1996): The number of types

Lexical Diversity Indices	Specifications and Quantification Methods
NDW-50	The number of types in the first 50 words of a text
NDWERZ /ER50	Malvern et al., (2004): Means of NDW for 10 random sub-samples of 50 words
NDWESZ /ES50	Malvern et al., (2004):Means of NDW for 10 random sub-samples of 50 consecutive words with random starting points

-T and V are used in this table to denote word types, N stands for the word tokens, Log stands for the logarithm, X is a vector with the frequencies of each type, fx is the frequencies for each x.
-Some of the measures indicated in this table are used for studies on language disorder, language impairment, and Alzheimer's.
-In brunet's W index, -a is a scaling constant that is set to – 0.172 in Tweedie and Baayen (1998); the lower values of W denote more diverse vocabulary. This value ranges between 10 and 20.
-Lower values of the Maas index denote greater lexical diversity.
-Hapax Dislegomena is the word types that only occur twice in a text.

Another ratio-based measure, the token-type ratio (also known as MWF or mean word frequency) which is the reciprocal of the well-known TTR, was used in the late twentieth century by Goldfield (1993) to examine the average number of nouns and verbs in the maternal speech to one-year-olds.

Apart from these quantification methods, earlier researchers used different equations which are simple manipulations of the original TTR method. In the earliest study, Linnarud (1975) uses the lexicality of both types and tokens for Lexical Variation. Some years later, in 1981, Mendelsohn favours only the lexicality of types, and Arnaud (1984) omits the percentage (refer to table 2.1).

There are a few other measures of diversity that are far-less studied and examined in the literature. One is the ID index or inflectional diversity (Richards & Malvern, 2004) which calculates the difference between the lexical diversity of inflected forms and stem forms and is reported to be sensitive to the number and variety of stems and inflections in the text.

Since all the above quantifying methods have been subject to various criticisms (particularly the sensitivity to text length), and none gives the perfect picture of lexical diversity, some researchers such as McCarthy and Jarvis (2010) recommend using a combination of them rather than a single index in research design, reminding researchers that lexical diversity can be assessed in various ways/indices, each gauging lexical variation from different angles.

2.2.3. Lexical Sophistication

A review of the literature on lexical complexity constructs points to several approaches to defining and quantifying lexical sophistication. Most researchers associate it with the use and/or percentage of rare words and/or less-frequently used or advanced vocabulary (Laufer & Nation, 1995; Read, 2000; Vermeer, 2004; Bardel & Gudmundson, 2012). The assumption is that students learn words roughly based on their frequency of occurrence (Kyle & Crossley, 2015; Nation, 1990; 1984; Sternberg & Powell, 1983; Vermeer, 2004), i.e., high frequency and general words are learned at early stages of language acquisition, processed more quickly and used more often; therefore, the presence of low-frequency words (sometimes referred to as 'advanced' words) in learners' productions indicates higher proficiency levels (g., Crossley et al., 2014; Kyle & Crossley, 2015; Perfetti, 1985; Rayner & Pollastsek, 1994; Sternberg & Powell, 1983). This, in turn, is based on the finding of Zipf (1932, 1935) who observed that the frequency of each word is inversely proportional to its rank (e.g., in a rank-ordered word list) and that a small number of words occur more often (i.e., few high-frequency words) and a larger number of words occur less-frequently in any natural corpus. This might account for the observation that L2 learners (especially low-proficiency ones) have a small lexicon that is mainly high-frequency words, while proficient learners have an extended lexicon which includes a lot of low-frequency and advanced words (Kyle & Crossley, 2015; van Hout & Vermeer, 2007). Consequently, frequency-based methods are used in most studies to measure lexical sophistication (Bardel & Gudmundson, 2012; Crossley & McNamara, 2013; Daller, Van Hout, & Treffers-Daller, 2003; Šišková, 2012; Vermeer, 2000; Waldvogel, 2014).

A well-accepted and widely-used quantification method for lexical sophistication is the use of word frequency bands (e.g., Laufer, 1994; Laufer & Nation, 1995; Morris & Cobb, 2004). A notable example is the Lexical Frequency Profiler (LFP; although they referred to it as lexical richness) that was designed by Laufer and Nation (1995) to examine the frequency bands in a text; LFP allocates all of a text's words into four frequency bands by reference to the word lists prepared by Nation (1984); it then profiles the proportion of word types in each band. To quantify lexical sophistication, they calculate the percentage of advanced tokens to the total number of lexical tokens. There is, however, little consensus on the definition of 'advanced' vocabulary, and as Laufer and Nation comment, the lack of a standard definition for the term 'advanced' causes complications since it depends on the learners' levels, educational system and the amount of instruction. Bardel and Gudmundson (2012), however, use the ratio of low-frequency to high-frequency words in productions. This allows assigning each learner a proficiency band based on native speakers' production corpora. Lindqvist, Gudmundson, and Bardel (2013) also use the frequency bands technique to measure lexical sophistication but they use the terms ‘lexical sophistication’ and ‘lexical richness’ interchangeably throughout their work. In 2.2.4 I will further discuss the terms that have been used interchangeably by various scholars. Their definition of lexical sophistication, however, is the infrequent words without cognates and thematic words. Similarly, Lindqvist et al., (2013) use frequency bands and the lexical profiling technique to develop the Lexical Oral Production Profile (LOPP) to measure lexical sophistication of oral data. They opted for using lemma as the counting unit rather than word family, arguing that it “reduces the number of forms attached to a headword” (p. 114), contrary to previous lexical profilers which were based on word families, such as Laufer and Nation's LFP (1995), and Cobb and Horst (2004, cited in Lindqvist et al., 2013). The LFP method and frequency bands were also implemented in the analysis of teacher talk (Meara, Lightbown, & Halter, 1997) and learner corpus (Bell, 2003).

Meara (2005b), however, critically analysed this tool and raised concerns about handling the errors, proper nouns, formulaic sequences, etc, and suggested an in-depth evaluation and modification of it before it becomes an established analytical tool. Nonetheless, it seems that the errors are corrected before the texts are processed by LFP. Formulaic sequences have not been measured by any automatic tool yet (Coh-Metrix analyser by Graesser, McNamara, Louwerse, and Cai, 2004, for instance, measures n-grams which is not the same thing as formulaic sequences). Furthermore, in Compleat Web VP (also known as VocabProfile, a web-based implementation of LFP based on the updated BNC/COCA frequency lists; v. 2.1, Cobb, 2019) there are three options for handling proper nouns: they can be ignored and classed as offlist, they can be eliminated from the text, or can be classed in the 1K words (the most-frequently-used and general words). Cobb (ibid.) favours the third option, arguing that the offlist category includes the rare words and assigning the proper nouns in this list makes a text with many proper nouns appear like a difficult text. He further raised concerns about eliminating the proper nouns as this compromises the density of known-to-unknown words in a text. Inclusion of the proper nouns in the 1K list, on the other hand, indicates that they are most likely to be known by many learners and that these words are interpretable in the context and do not impose learning burden.

Another method based on word frequencies is the use of a corpus or multiple corpora and their derived word lists (usually high-frequency word lists) as external reference points to judge the sophisticated items as those that do not appear among the high-frequency words in these lists (Lu, 2012; McNamara, Crossley, & McCarthy, 2010). For instance, McNamara et al. (2010) opted for measuring lexical sophistication as less-frequently-used words based on the CELEX (Baayen, Piepenbrock, & Gulikers, 1995) word frequency; CELEX is a database of frequencies from an early version of COBUILD (Collins Birmingham University International Language Database) corpus. They showed that the use of less-frequency words suggests higher lexical proficiency. Similarly, the word frequency indices in Coh-Metrix uses the CELEX database as the baseline. The lexical sophistication indices in LCA (Lexical Complexity Analyzer, Lu, 2012) are also judged as those that do not appear in the top 2000 most-frequently-used words in the BNC (British National Corpus) word list with an alternative to use the ANC (American National Corpus) word list. Similarly, its modified version, LCA-AW (Lexical Complexity Analyzer for Academic Writing, Nasseri & Lu, 2019) specifies lexical sophistication indices as the words that do not appear among the 2000 high-frequency words in the BNC or ANC nor appear in the BAWE (British Academic Writing English) corpus' word list for linguistics and language studies. The British Academic Written English Corpus, was an ESRC project that was carried out by Hilary Nesi, Sheena Gardener, Siân Alsop, Paul Thompson, Paul Wickens, Maria Leedham, and Signe Oksefjell Ebeling from 2004 to 2007. Lexical sophistication Indices in these two analysers include LS1 (lexical sophistication type I, calculated as N_{slex} / N_{lex}), LS2 (lexical sophistication type II, calculated as T_s / T), VS1 (Verb Sophistication type I, calculated as T_{sverb} / N_{verb}), VS2 (Verb Sophistication Type II, calculated as T^2_{sverb} / N_{verb}), and CVS1 (Corrected VS1, calculated as $T_{sverb} / \sqrt{2}\, N_{verb}$), where N is the number of tokens, T is the number of types, lex stands for lexical, s stands for sophisticated, and sverb stands for sophisticated verbs. Apart from LS2 which uses all types (unique words, both lexical and function words), the rest of the lexical sophistication measures use lexical or content words (types and tokens). More details about LCA-AW as a contribution of this study will be presented in 5.3.2 and the way to access it and analyse the texts will be explained in Appendix D. Crossley, et al. (2013) as well as Kyle & Crossley (2015) also agree with this latter frequency-based approach, arguing that this approach produces more accurate predictor models than the frequency-band approach.

Kyle and Crossley (2015, 2016), however, regard lexical sophistication as an embodiment of the indices of lexical frequency, range, n-gram frequency, academic vocabulary, and word information properties as calculated by TAALES (Tool for the Automatic Analysis of Lexical Sophistication, Kyle & Crossley, 2015). The academic lists for frequency counts in TAALES are not discipline-specific (i.e., they consist of several different fields of study), while in LCA-AW, the academic word list (i.e., frequently-used academic words) is specific to linguistics-related disciplines derived from the BAWE corpus, and hence more suitable for analysing texts in linguistics and language studies.

Another instance of the use of the term ‘lexical sophistication’ is in Daller and Xue (2009) who consider lexical sophistication as an umbrella term for two measures of Lexical Frequency Profile and Guiraud Advanced. As explained in the previous section, however, some researchers consider the Guiraud Advanced (also called Advanced Guiraud) as an index of lexical diversity and a representative of varied and non-repetitive vocabulary, unless they specify ‘advanced’ as sophisticated or less-frequently-used words, e.g., based on a reference word list.

Although not named specifically by the term ‘sophistication’, the usage of rare words in early vocabulary composition is quantified based on the type-type ratios, for example, noun types per verb types (see the related discussions in Linnarud, 1983 and Malvern et al., 2004). This use of rare words as representative of sophisticated vocabulary can also be seen in Arnaud’s (1984) ‘score of rareness’; he considers the proportion of rare types to lexical types, instead of total tokens ($R = V \text{ rare} / V_{\text{lex}}$) to measure lexical sophistication. Along this line, a measure of ‘rare word density’ (also called exposure to rare words) was used by Snow, Tabors, and Dickinson (2001) in their home-school study of language and literacy development. This type-type index calculated the proportion of word types in their transcript that were judged to be rare. There are two other less-frequently-studied measures of lexical sophistication as the use of rare words, namely hapax legomena (i.e., the word types that only occur once in a text) and hapax dislegomena (types that only occur twice in a text). These two measures are either calculated as simple frequency counts, or as a proportion. This leads us to the concept and/or measure of Honoré’s statistic (also represented as R or H for brevity; Honoré, 1979, cited in Holmes & Singh, 1996) which is mainly used in stylometric text analysis. It is calculated as the proportion of hapax legomena in a text ($R = 100 \log (N) / (1 – v1 / V)$ where v1 is the number of hapax legomena. This measure, which in Malvern et al. (2004) is classed as lexical diversity, is also used in the studies on Alzheimer’s and aphasic learners (for detailed discussions on this topic see Malvern et al., 2004 as well as Sichel, 1986). The proportion of rare tokens (e.g., in Dickinson, 2001) is perhaps a rare use of this type of lexical sophistication.

There are a number of *ad hoc* definitions for the term lexical sophistication in the SLA and corpus-based studies as well. For instance, lexical sophistication is defined by Dascălu, Trausan-Matu, & Dessus (2012) as “the complexity of a word’s form in terms of the average number of characters” (p. 272). This is based on the interesting observation by Zipf (1932) that word length is inversely proportionate to the frequency of usage (e.g., shorter words are used more often than longer ones) and that longer words denote higher lexical proficiency. Another *ad hoc* use of lexical sophistication is the measure of ‘lexical originality’ (also called ‘lexical individuality’, Read, 2000) that calculates the percentage of words exclusively used by one writer compared to other writers in a corpus. Linnarud (1983), for example, divides the writer-specific words to the total number of lexical words. Malvern et al., (2004) however, raise concerns about this measure’s lack of specificity regarding its ratio form, e.g., whether it is implemented as a type-token or token-token ratio.

Some scholars regard lexical sophistication as the representation of width of vocabulary knowledge similar to what was specified earlier as lexical diversity (e.g., Housen et al., 2008) and some as the depth and breadth of lexical knowledge (e.g., Kyle & Crossley, 2015; Read, 1998; Meara, 1996, 2005a). The former case will be further explained in the next section. The two terms of ‘lexical knowledge’ and ‘vocabulary knowledge’, however, seem to be used interchangeably in the literature. Even though in this study I clarified the measures based on all words or word types vs. lexical (i.e., content words) types and tokens in table 5.3. and section 5.3.2, I continue to use these two terms interchangeably because the analysers that are used in this study treat these categories as ‘lexical’.

I also argue that ‘rare’ and ‘less-frequently-used’ words are highly context-dependent because a word that is used infrequently in one corpus may, in fact, be quite frequently used in another corpus/text. Instances of such words are discipline-specific terminology in a related corpus. It seems more plausible, therefore, to screen the sophisticated words of a text from a specific genre or discipline based on the frequently-used words derived from that discipline.

2.2.4. The Cases of Mismatch between Terms, Definitions, and Measurement Criteria of Lexical complexity constructs and measures

Even though the main body of literature attests to the definitions/specifications of the three constructs of lexical density, diversity, and sophistication as elaborated in the preceding sections, there are a considerable number of inconsistencies regarding the use of the terms, definitions and quantification methods of these constructs and other terms such as lexical richness. I emphasise that some of these cases of inconsistencies are a simple mismatch between the preferred use of such terms, while others seem not to fit entirely in any classification.

This inconsistent use of the terms of lexical proficiency and an absence of a unified position on the distinction between the terms that are similar and/or used interchangeably has already been noticed by some researchers (e.g., Malvern et al., 2004; McCarthy & Jarvis, 2010). For example, both McCarthy and Jarvis (2010) and Malvern et al., 2004 notice the absence of a unified position as to whether we should distinguish between lexical diversity, vocabulary diversity, and lexical richness. By providing a review of such inconsistencies, it is hoped that future researchers new to this field can navigate through the studies and make informed decisions about the use of these terms and the selection of certain lexical indices. In what follows, I judge these three constructs based on the theoretical and conceptual understanding of them (see the detailed explanations in 2.2.5) as well as the most common use of these terms, definitions, and measurement criteria in the literature. In this regard, lexical density is the proportion of lexical items in a text, lexical diversity is the use of varied/diverse and non-repetitious words (also known as unique word types), and lexical sophistication is the proportion of advanced vocabulary and/or less-frequently-used words filtered through the most-frequently-used word lists in different corpora or based on frequency bands.

As elaborated in 2.2., the two terms of lexical richness and lexical complexity have been often used interchangeably to denote a set of constructs (and their respective indices) of either or all of the constructs of lexical density, diversity, and sophistication.

There are several studies which equate lexical diversity with lexical richness in terms of measurement, for example, type-token ratio and its many mathematical and computational variants proposed (e.g., Arnaud, 1984; Stajner & Mitkov, 2012; Tweedie & Baayen, 1998; van Gijsel, Speelman, & Geeraerts, 2006; Vermeer, 2000; and Wimmer & Altmann, 1999). Lexical richness has also been used interchangeably with lexical sophistication (Bardel & Gudmundson, 2012; Lindqvist et al., 2013); in some studies, it is used as an umbrella term for lexical diversity and density (Linnarud, 1975), or even equated with lexical complexity (i.e., the definition of lexical complexity in this thesis) to encapsulate a range of lexical measures (Laufer & Nation, 1995; Read, 2000; Lu, 2012; Kim, 2014). Consequently, there are three main approaches to understanding and measuring lexical richness. One uses the same technique as measuring lexical diversity, namely the type-token ratio and its many variants proposed for examining the use of varied and non-repetitious vocabulary. The second which is sometimes used interchangeably with lexical sophistication as well, is measuring a text's vocabulary richness based on a set of word frequency bands as an external reference point, such as LFP (e.g., in Daller & Xue, 2009). In the third, it is used as an umbrella term for various lexical constructs and measures (e.g., Kim, 2014; Lu, 2012; Read, 2000). Some of the above studies analysed one or some other linguistic features including syntactic complexity, lexical fluency, grammatical accuracy, and proportion of errors along with these lexical measures as well (Read, 2000; Schcolnik, 2008; Johnson et al., 2012; Vaezi & Kafshgar, 2012; Mazgutova & Kormos, 2015).

Other instances of mismatch between the terms, specifications and quantification methods include the use of LFP in Laufer and Nation (1995) to measure ‘lexical richness’, while the same technique is used in Bardel and Gudmundson (2012) to measure ‘lexical sophistication’; this most probably is a simple mismatch between the terms only. Besides, Laufer and Nation first list various measures such as lexical originality, lexical density, lexical variation and lexical sophistication under the umbrella term 'lexical richness'; however, they conclude that none of these measures could effectively capture lexical proficiency of a learner and consequently they offered LFP for measuring ‘lexical richness’. It is not entirely clear whether they still regard LFP as an alternative which can reflect all the mentioned measures or as a separate entity which only focuses on the type and rarity of words based on the frequency bands.

Furthermore, for Housen et al. (2008) ‘lexical sophistication’ is featured as the knowledge of semantic relations and fits in the macro-level of the lexicon, associated with lexical width. They define this measure as the learner's knowledge of “different but related lexical alternatives for referring to a referent” (p. 3). They further use the term ‘lexical sophistication’ being conceptually devised as “semantically more specific and/or pragmatically more appropriate different words” which correlates with the knowledge of semantic relations such as synonymy, antonymy, hypernymy and hyponymy (Housen et al., 2008, p. 3), what most researchers would refer to as 'lexical diversity', i.e., the use of varied and different words.

The only issue arising from such cases of mismatch between the terms, definitions, and quantification methods is the difficulty in interpreting and comparing various works with various measurement criteria and the possibility of misreading the results, especially by novice researchers. In the absence of a unified framework of analysis and consensus on the measurement criteria, it seems indispensable that each researcher should clarify the exact criteria while discussing and interpreting the results of the works with different measurement criteria or quantification methods.

2.2.5. Construct-distinctiveness of Lexical Density, Diversity, and Sophistication: Theoretical and Conceptual Perspectives

Some researchers view lexical density as an aspect of a text's lexical diversity (Stamatatos, Fakotakis, & Kokkinakis, 2000; Stajner & Mitkov, 2012). Here I offer a rationale for considering these two measures as separate entities which carry different implications when describing a text.

Lexical diversity and density, although interrelated, can be differentiated in that lexical density seeks to present how densely lexical items are packed into syntactic structures, while lexical diversity is representative of non-repetitious and/or different lexical and grammatical items. Lexical density as such can accommodate morphological variants of a lemma (Stajner & Mitkov, 2012), and is tightly related to the knowledge of syntactic structures which can carry those morphological variants, while the knowledge of these morphological variants (inflections, compounds, and derivations) does not necessarily represent a diverse knowledge of vocabulary and therefore may be dealt with separately rather than being accounted for in lexical diversity/variation formulas. Correspondingly, a learner can produce statements with higher lexical density and lower lexical diversity and vice versa (Johansson, 2008). Linnarud (1975) also confirms that the results of some studies testify to high values of lexical density with poor and repetitive vocabulary (low lexical diversity) of the same texts. As an instance, the process of nominalisation reduces the grammatical words and contributes to higher lexical density. Consequently, the texts with these characteristics are more informative and can be regarded as a characteristic of the academic genre and advanced writing (Biber, 1988, 2006; Biber, Gray, & Poonpon, 2011; Ryshina-Pankova, 2015), while a high value in lexical diversity can also be achieved by a beginner or intermediate learner who knows how to use all the limited supply of vocabulary diversely.

By the same token, the term 'diversity' signifies variety and lexical diversity seeks to demonstrate the use of diverse and non-repetitious words used in production which are not necessarily advanced words (in terms of the rarity of occurrence or being less-frequently used). Therefore, it is possible to have a high lexical diversity value in a text, but low lexical sophistication. Consequently, quantification methods need to become distinct rather than using the same concept of TTR. Despite the overlapping areas between the concepts of lexical diversity and lexical sophistication regarding the use of high versus low-frequency words, the two constructs do not necessarily correspond. I will demonstrate how a learner can achieve higher lexical diversity without necessarily using/knowing advanced or low-frequency words. Lexical diversity correlates with the knowledge of synonyms and learners can use diverse and non-repetitious word types for the same concept while not resorting to low-frequency and advanced words as distinguished based on word frequency bands. Take the following example of words with similar literal and/or conceptual meanings: say, tell, state, present, declare, remark, mention, assert, utter, speak, express, indicate, articulate, postulate, pronounce, vocalize, talk, verbalize. Among them, those which appear in the first 1000 most frequently-used words (MFUW) in the COCA list, for example, are: ‘say’, ‘tell’, ‘talk’, ‘state’, ‘present’, ‘speak’, ‘express’, ‘indicate’; those which appear in the second 1000 MFUW are ‘declare and express’; ‘assert, remark, and articulate’ appear in the 3000-5000 MFUW, and the rest of the verbs namely, ‘utter, postulate, pronounce, vocalise, and verbalise’ do not appear in the top 5000 MFUW at all. Therefore, the low-frequency words, such as those which appear in the 2000-5000 word list are classed as advanced words based on the definition of lexical sophistication in lexical profilers. A learner may use all or most of the eight high frequency words which contribute to the overall lexical diversity of a text without attempting any of the low-frequency or advanced words which indicate a sophisticated text. On the same ground, and theoretically speaking, a learner can use two words of 'postulate' and 'articulate' which belong to sophisticated words, repeatedly and contribute to lower overall lexical diversity while still exhibiting advanced vocabulary. In practice, however, a learner with the knowledge of advanced words is more likely to use a wider range of vocabulary and hence higher lexical diversity as well, but the opposite is not necessarily true as mentioned above.

In chapter six, I will further examine the construct-distinctiveness of lexical density, diversity, and sophistication based on the academic writing corpus in this study via correlation and factor analyses.

2.3. Syntactic Complexity Constructs and Measures: Terms, Definitions/Specifications, and Quantification Methods

The word 'syntax' designates principles of the grammatical arrangement of words and morphemes in phrases and sentences to form meaningful combinations. Syntactic complexity, as a result, refers to the range, type, and complexity of syntactic structures, often quantifiable via measures such as the number of words per T-unit, the mean length of sentences, dependent clauses per clause, etc.

Syntactic complexity which in Ortega (2003) is equated with linguistic complexity and syntactic maturity, is defined as “the range of forms that surface in language production and the degree of sophistication of such forms” (p. 492) and in Lu (2014) is defined as “the range and degree of sophistication of syntactic structures” (p. 130). Likewise, Pallotti (2015) states that the complexity of syntactic structures depends on “the number of constituents and the number of combinations they may take” (p. 123); even so, he poses that complexity of a certain syntactic pattern is often described theoretically than being grounded in research.

There is a rapidly growing literature on the applications of measures of syntactic complexity in various language-related research such as child language acquisition, language impairment and readability formulas. However, specific attention has been placed on second language writing research to examine the roles of syllabus design, assessment, task complexity and explicit teaching on grammatical and structural development and writing ability, as well as examining L2 learners' texts or oral productions concerning variables such as age, proficiency level, gender and timescale.) Syntactic complexity is regarded as reliable measured aspects of writing ability (Rafoth and Combs, 1983) and its representative measures as indices of language development and proficiency (Bulté and Housen, 2014). Likewise, Beers and Nagy (2009) view it as a predictor of adolescent writing quality and believe that certain complex structures like the amount of embedding help with the expression of complex ideas and concepts and the elaborate relationships among such concepts.

Various constructs proposed in the studies such as “length of production unit, amount of embedding, range of structural types, and sophistication of the particular structures” (Ortega, 2003, p. 492) are quantified via their representative measures such as the length of T-units, which is in turn derived from the frequency of its base production unit, e.g., T-units. With language development in mind, she advises that measurement methods of syntactic complexity “have to strike a balance between reliability, feasibility, and sensitivity to language development theory” while being “reasonably easy to calculate” (Ortega, 2000, p. 4). She further advises that syntactic complexity measures might need to be revisited in order to be employed for both written and spoken discourse as each mode may exhibit its own peculiarities such as less structurally complex sentences in spoken discourse. The effect of modality and its relationship with syntactic complexity as well as the usefulness of certain syntactic indices in each language mode have also been addressed in Larsen-Freeman (1983), Biber (1988), and Halliday (1987, 1989).

During the past two decades, syntactic complexity studies have has witnessed a growing number of proposed measures as indices of L2 proficiency. In the following sections, I review the most-frequently-used and reported syntactic structures and measures in first and second language acquisition and development, L2 performance and proficiency (differences), and writing research. Although some researchers have used the terms 'syntactic' and 'grammatical' complexity and/or structures interchangeably (e.g., Bulté & Housen, 2012), in this thesis I distinguish between the two in that syntactic complexity includes overall constructs and their constituent measures which reflect structures more than a word, for example, phrases, clauses, T-units while grammatical complexity includes fine-grained measures, usually at the word level. In reporting the syntactic measures and constructs, however, I also include studies that use the term 'grammatical' complexity to refer to the overall structures as defined in this thesis. For example, Biber and Gray (2010) as well as Biber, Gray, and Poonpon (2011) used the term 'grammatical complexity' to refer to two types of measures as standardised rates of occurrence of specific structures.

Both Wolfe-Quintero et al. (1998) and Ortega (2003) acknowledge the conflicting results as well as the mismatch between the terms and definitions in various studies, and attribute these to the variations in research design, especially the task types, sample/corpus, and the operationalisation of proficiency. Lu (2010, 2011) therefore emphasises the necessity to explicitly define the terms and definitions for each analytical unit and to specify the measurement criteria and quantification methods.

In what follows, I present five important and overall syntactic constructs along with the indices that quantitatively represent these constructs. Since indices as indicators and/or predictors of proficiency differences in advanced levels in the context of academia are infrequently reported in the literature, I will include developmental indices in SLA and writing research as well, with a focus on the measures that are recommended by Lu (2010, 2011) for analysing written productions of advanced L2 learners, as computed in L2SCA (L2 Syntactic Complexity Analyzer, Lu, 2010).

2.3.1. Length of Production Units

Among the various proposed syntactic structures, measures pertaining to the construct of 'Length of Production', such as the average number of words per T-unit, sentence, or clause have had a longer shelf life (e.g., in Crowhurst, 1983; Golub & Frederick, 1971; Lu, 2010; Lu & Ai, 2015; Mancilla, Polat, & Akcay, 2015; Ortega, 2003; Witte & Davis, 1982; Wolfe-Quintero, et al., 1998). For example, Wolfe-Quintero et al. (1998) who reviewed thirty-nine studies on second language writing concluded that several metrics such as mean length of clause and mean length of T-unit are good indicators of linguistic proficiency levels. In her research synthesis of twenty-one cross-sectional and six longitudinal studies (EFL and FL groups), Ortega (2003) also investigated the significance of syntactic complexity measures on the proficiency of college-level L2 writers. Common among the studies were most frequently-used measures of Mean Length of Sentence (MLS), Mean Length of T-unit (MLTU also abbreviated as MLT), and Mean Length of Clause (MLC). The same measures are also selected by Lu’s (2010) study of second language writing as well as Lu and Ai’s (2015) investigation of syntactic complexity in college-level English writing. Similarly, according to Ortega (2000), in most L2 and L1 studies, length of production units (e.g., mean length of T-unit) together with the amount of embedding are among the main investigated metrics. The significance of length-based syntactic structures could be partly attributed to Brown (1973, cited in Ortega, 2003) who initially recommended including “mean length of utterance” (MLU) in studies of child language development.

Hunt (1965) originally introduced the T-unit - which is a minimal terminal unit and includes one independent clause plus any dependent clauses - as a criterion in measuring sentence development in school children's writing. As children tend to produce more run-on than complete sentences, each of their sentences includes several T-units, which again tend to be longer as they get older. In 1970, Hunt refines this definition to “a main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it” (p. 4). A complex T-unit, consequently, includes at least one independent clause and at least one (usually more than one) dependent clause (Casanave, 1994; Lu, 2010). Golub and Frederick (1971) refer to this as multi-clause T-units, instead of complex T-units. The difference between a simple T-unit and a complex one is that a simple T-unit requires only one independent clause and the dependent clauses are optional, while a complex T-unit requires at least one independent and one dependent clause.

Length of T-unit which interestingly enough was sometimes considered as indices of lexical richness with syntactic properties, also caught the attention of Laufer and Nation (1995). In 1978, Larsen-Freeman described the design of an index of ESL development for prospective EAP students based on the written placement exams. The values of the syntactic measures of ‘average words per composition’ and ‘average words per T-unit’ were shown to significantly differ between proficiency groups.

Other studies commented on the T-unit's overdependence on subordination and discounting the complexifications arose by coordination (Bardovi-Harlig, 1992; Ortega, 2000) and therefore, as they argue, it is practically less effective for analysing spoken discourse with natural features such as ellipsis and more suitable for analysing written discourse. T-unit analysis, as they comment, breaks up coordinated sentences - especially those with additive function - as well as asymmetrical conjunctions, such as conjunctive-conditionals, and thus ignores the grammatical and rhetorical sophistication they carry. In the latter case, T-unit analysis breaks up semantic and syntactic units and renders conjunctions “as semantically null” (Bardovi-Harlig, 1992, p. 392). Kyle (2016), however, argues that MLT (compared to MLC, as is discussed below), “adds an extra level of specificity” in that dependent clauses are disambiguated (e.g., because they are attached to the independent clause as a whole unit).

Mean length of clause (MLC) or the average number of words per clause is viewed as a “global measure of intra-clausal complexity” (Kyle, 2016, p.9). A clause is specified as a structure with a subject and a finite verb (Hunt, 1965; Lu, 2010; Polio, 1997); some other researchers such as Bardovi-Harlig and Bofman (1989) include non-finite verbs in the definition of a clause. Clauses include nominal clauses, adjective and adverb clauses as well as independent clauses. The MLC index is also viewed as an indicator of linguistic proficiency (Kyle, 2016; Lu, 2010; Wolfe-Quintero et al., 1998). Kyle (2016) argues that the value of MLC is affected and increased by an increase in phrasal coordination, the use of perfect and progressive aspects (because they require two auxiliaries compared to other aspects), and longer syntax structure types, e.g., SVO compared to the simple SV syntax type. He further emphasises that MLC is an overall clausal complexity level since it does not differentiate between dependent and independent clauses and both types are considered on an equal footing.

Mean length of sentence or MLS is another length-based index which is less-frequently investigated compared to MLC and MLT (e.g., in Alexopoulou et al., 2017). MLS, however, has an advantage over the other two indices because of its relatively straightforward definition and operationalisation. A sentence is commonly calculated as a string of words that start with an uppercase letter and end with one of the end-of-the-line punctuations of period, question mark, exclamation point, or ellipsis (Hunt, 1965; Kyle, 2016; Lu, 2010). MLT and MLS are shown to have a high positive correlation in Lu (2010) while a sentence can contain multiple T-units. The MLS index is also shown to have a positive relationship with language proficiency (see for instance the results of Alexopoulou et al., 2017 and the studies reviewed by Wolfe-Quintero et al., 1998 and Ortega, 2003).

2.3.2. Subordination Structures and Indices

There is a wealth of studies that consider subordination as a distinct syntactic construct and as a characteristic of L2 production complexity (e.g., Alexopoulou et al., 2017; Homburg, 1984; Ortega, 2000) while Wolfe-Quintero et al. (1998) concluded that certain subordination indices (such as C/T as will be discussed) are more related to proficiency based on programme and school level rather than the holistic ratings or short-term changes. Various subordination indices are reported to be indicators and discriminators of proficiency (see for example the discussion and findings in Grant & Ginther, 2000; Kim, 2014; Lu, 2010, 2011; Ortega, 2003).

Both dependent and independent clauses are predominantly used as structures comprising the subordination indices (e.g., Homburg, 1984; Kyle, 2016; Lu, 2010, 2011, 2014; Ortega, 2000, 2003; Wolfe-Quintero et al., 1998 among many others). A clause (e.g., independent clauses, nominal, adjective and adverb clauses) is defined as a structure comprising a subject and a finite verb (Hunt, 1965; Lu, 2010, 2014; Polio, 1997) and a dependent clause is specified as a finite nominal, adjective, or adverb clause (Cooper, 1976; Hunt, 1965; Lu, 2010, 2014) whose meaning is not complete.

Four main subordination indices that are reported in the literature and are computed in L2SCA are C/T or T-unit complexity ratio, CT/T which is a complex T-unit ratio, DC/C that is a dependent clause ratio, and DC/T which measures dependent clauses per T-unit. Since these four indices are investigated in this study, they will be described in detail in 5.3.1.2.

A similar index to C/T is ‘the number of clauses per sentence’ which together with the indices of ‘the number of clauses per main clause’ and ‘the average value of embedded clauses’, that was proposed by Arena (1982, cited in Kyle, 2011), was investigated in Sparks’s (1988) study of ESL academic writing. All three measures were reported to be reliable measures based on holistic ratings. A similar index to DC/C is also IC/C measures ‘independent clauses per clause’ and is investigated in Ortega (2000) but is listed as a coordination ratio and will be discussed in the next section.

The frequency of ‘Subordinate noun clauses’ was also taken as an index in Golub and Frederick’s (1971) study of linguistic structures of students in upper elementary grades. They define this measure as “a clause occurring in one of the functions common to a noun (subject or object of a verb, object of a proposition)” (p. 12). Additional related subordination structures in their study were ‘subordinate adjective clauses’, defined as “a clause modifying a noun or a word used as a noun”, as well as ‘subordinate adverbial clauses’ which they specified as “a clause which functions as an adverb, i.e., it modifies a verb, a verbal, an adjective, an adverb, or another clause". They further investigated a rather vague index called 'other subordinate clauses' which include any type of dependent clauses that do not function as a noun, adjective or adverb clause; for example, any clause following the expressions of 'looks like', or 'seems like' (p. 12).

An extended set of 15 fine-grained subordination indices were also used in Ortega (2000) in three broad categories of noun clauses, relative clauses, and adverbial clauses, each with five distinct and fine-grained indices. The noun clause category consists of the five measures of 'noun clauses per sentence' or Noun/S, 'noun clauses per utterance' or Noun/U, 'noun clauses per T-unit' or Noun/TU, 'noun clauses per clause' or Noun/C, and 'noun clauses per dependent clause' or Noun/DC. The second category consists five indices with relative clauses as the numerator and sentence, utterance, T-unit, clause, and dependent clause as denominators: Rel/S, Rel/U, Rel/TU, Rel/C, and Rel/DC. The third category works in similar ways and consists five indices with adverbial clauses as the numerator and the same five production units as denominator: Adv/S, Adv/U, Adv/TU, Adv/C, and Adv/DC. She argues that these fine-grained subordination measures have rarely been investigated in SLA and writing research and therefore there is little known about their predictive power in L2 discourse. Among the few such studies, Cooper (1976) examined the amount of Adv/TU or the adverbial subordination per T-unit in L2 German production but found insignificant differences across programme levels. Kameen's (1979) investigation of noun, relative, and adverbial clause frequencies as well as Sharma's (1980) research on relative clause production were other instances that found a relationship between the increased values of these measures and higher levels of writing ability; they, however, did not find any straightforward relationship between the values of these indices and holistic ratings.

An interesting index of embedding depth was also studied by Salah (1990) to test the hypothesis that "clause is the primary unit of information" (p. 121). The concept behind this index is the idea of clause depth based on the standard transformational theory. According to this theory, the embedded clauses in a sentence are processed one clause at a time, "starting with the lowest clause, followed by the next higher clause cycling upward until the main clause is reached" (p. 122). This, in turn, affects the processing time, and hence a deeper and more embedded structure is believed to be more complex. To get a value for this index, all clauses in a discourse need to be separated based on type and frequency and numerical values are assigned to each based on the clause analysis scheme in Salah. Ortega (2000, p.26) commented on this type of clause analysis that these measures are "very laborious and require extensive training".

Three measures related to the above index measuring ‘depth of clause’ are also investigated in Ortega (2000) as ‘clauses per sentence’ or C/S, ‘clause per utterance’ or C/U, and ‘clauses per T-unit’ or C/T. The latter, which was discussed earlier, is used by Lu (2010, 2011) as well as Lu and Ai (2015). Ortega (2000) also argues that depth of clause ratios have an advantage over length-based ratios in that “observed increases in length of production unit on these measures can only be attributed to clausal elaboration” (p. 40). She further emphasises that the indices based on the depth of clause and the subordination measures, in fact, gauge syntactic complexity in similar ways, that is the amount of elaboration (e.g., clausal elaboration) via subordination.

2.3.3. Coordination Structures and Indices

The use of coordination structures is believed to be a characteristic of syntactic complexity in early L2 development (e.g., Ortega, 2000; Sato, 1990; Wolfe-Quintero et al., 1998) in that the increase in coordination is marked as a developmental stage in L2 writing complexification (Bardovi-Harlig & Bofman, 1989; Wolfe-Quintero et al., 1998). This point will be further discussed in the next chapter. Coordination structures include coordinate phrases (CP), coordinate clauses, and sentence-level coordination. Ortega (2000) also documents how an increase in subordination leads to a decrease in coordination. Coordinate phrases, for instance, coordinates/conjoins more than one phrase including noun, verb, adjective and adverb phrases (Cooper, 1976; Lu, 2010, 2014) using coordinating conjunctions (e.g., ‘and’, ‘but’, ‘yet’, ‘both … and’, ‘neither … nor’, ‘whether … or’).

Three main coordination indices that are reported and computed in L2SCA are ‘coordinate phrases per clause’ (CP/C), ‘coordinate phrases per T-unit’ (CP/T), and ‘sentence coordination ratio’ (T/S). The first two of these indices will be investigated in this study and a detailed review of them in the literature will be presented in section 5.3.1.2. The T/S measure represents the ratio of the number of T-units to the number of sentences and is indexed as a sentence coordination ratio in Lu (2010) and measures the amount of independent clausal coordination. As Kyle (2016) explains, an index score of ‘2’ for instance, means that on average, every sentence in the analysed text includes one instance of clausal coordination. While Lu (2010) did not find any between-group differences regarding the values of this measure in academic writing proficiency studies, Monroe (1975) reported T/S index as an indicator of language proficiency and that clausal coordination decreased with the increase of proficiency.

The independent clause per clause index (IC/C) as a coordination ratio is also investigated in Ortega (2000). A “sentential-coordination” index is also proposed by Bulté and Housen (2012) which calculates the ratio of coordinate clauses to clauses, but has not been studied so far to the best of my knowledge. However, they did not elaborate on this index and its quantification method and it is not clear if this is the same index as the IC/C or independent clauses per clause index that is recommended by Wolfe-Quintero et al. (1998) and used in Ortega (2000) as a coordination ratio.

Another measure is the 'Coordination Index' which is defined as “the degree to which a learner achieves syntactic complexity through coordination” (Bardovi-Harlig, 1992, p. 393). It can roughly be illustrated as:

Coordination Index = [Independent-clause coordination / (clauses – sentences)] x 100

This index which was developed as an alternative to T/S (Sentence Coordination Ratio, as discussed earlier) differentiates between the amount of coordination and that of subordination. This index which is also investigated in Ortega (2000), differs substantially from Hunt's “main clause coordination index” (1970, p.189 as cited in Bardovi-Harlig, 1992), in that this formula takes into account “multiclausal sentences” and presents the coordination frequency relative to the number of combinations, while Hunt's index is a ratio of the sum of the number of T-units to the sum of the number of sentences (T/S).

Finally, the ‘coordinated T-units’ index was investigated in Golub and Frederick’s (1971) study of detecting linguistic structures of upper elementary grades. T-units were judged as coordinated if they were not separated by a period followed by capitalisation (e.g., the start of a new sentence).

2.3.4. Phrasal Complexity, Sophistication and Structures

Phrasal complexity and sophistication indices and structures have been infrequently used in first and second language acquisition and development, writing research, and studies on linguistic performance, proficiency and development as well as register variation studies. McNamara et al. (2010) for instance reported that phrasal-level syntactic complexity features are good distinguishers of L2 writing quality. Biber and Gray (2013), Biber, Gray, and Poonpon (2011), as well as Liu and Li (2016) equally recommend the investigation of phrasal-level structures like noun phrases and nominalised structures as distinct features of advanced academic writing.

Prominent among phrasal sophistication and complexity measures are the CN/C index which calculates complex nominals per Clause and CN/T which does the same in T-units. These two indices are selected and analysed in Kyle (2016) as well as Lu (2010, 2011) and Lu and Ai (2015). The two indices were shown to have a high positive correlation in Lu (2011) and Kyle (2016) with a correlation coefficient of above 0.8. Complex nominals based on the specifications of Cooper (1976) and Lu (2010) capture nominal clauses, gerunds and infinitives in subject position, as well as the nouns plus adjective, participle, appositive, prepositional phrase, and relative clause. VP/T or verb phrases per T-unit index is another important phrasal complexity measure that calculates the number of verb phrases in a T-unit and includes verb phrases with both finite and non-finite verbs. These three global indices of phrasal complexity will also be investigated in this study and therefore, described in more detail in section 5.3.1.2.

Apart from the discussed mean-based and ratio-based measures, the frequency of occurrence of certain phrasal structures and/ or the rate of their occurrence in a fixed number of words (per 100 or 1000 words) were also investigated; the latter indices are instances of standardised measures. Some instances of such indices are ‘appositive noun phrases as nominal post-modifiers’, ‘the amount of nominalisations’, and ‘rate of attributive adjectives per 1000 words’ (e.g., Biber & Gray, 2010, 2013; Biber, Gray, & Poonpon, 2011).

2.3.5. Other Indices and Analysis Approaches

Most of the syntactic indices and structures that have been discussed in previous sections, gauge mainly global-level syntactic complexity, for example via the mean-based and ratio-based measures that count the mean number of certain syntactic structures in a unit of production such as T-unit, clause, or sentence. These indices often calculate ratios as covered by the entirety of a text. There is also a fundamentally different approach to such analysis which is referred to as the standardised rates of occurrence of specific grammatical structures to operationalise grammatical and syntactic complexity. Biber and Gray (2010) as well as Biber, Gray, and Poonpon (2011) among other similar works, for instances, rely on such standardised measures to investigate important lexico-grammatical features in register variation studies. The examples of such indices are the rate of finite complement clauses per 1000 words and the rate of attributive adjectives per 1000 words. They refer to this approach as the register/functional approach (for detailed discussions see Biber, Gray, & Staples, 2016).

Sentence Complexity Ratio (C/S) is also an important measure that is considered in L2SCA classification of syntactic indices. This index is quantified as the ratio of the number of clauses to the number of sentences (Kyle, 2016; Lu, 2010; Lu & Ai, 2015) representing the overall sentence complexity. This is listed as a global index since it measures both the amount of clausal coordination and the amount of subordination in each sentence. This index has been reported to have a positive relationship with language development (Ishikawa, 1995) but a negative relationship with the school year (Lu, 2011).

There is also a mention of an Index of Complexity in Flahive and Snow (1980) in which each T-unit obtains a complexity score and the index is calculated as the ratio of this score by the number of words per T-unit. These complexity scores were in turn based on the frequency of certain grammatical structures. For example, adjectives and derivational morphemes were assigned a score of ‘1’, passive sentences, embedded questions, and relative clauses were given a score of ‘2’, and a score of ‘3’ was given to noun clauses. However, this index of complexity was not successful in discriminating between proficiency levels in their study. The complexity of T-units in the numerator of this index, however, should not be confused with the definition of complex T-unit in Lu (2010) that is specified as any T-unit which consists of at least one dependent clause.

There are a number of other less-frequently-used and reported syntactic indices such as the number of passive constructions per T-unit, per clause, and per sentence (e.g., the Kameen’s (1979) study reviewed in Wolfe Quintero et al., 1998). These three indices were reported by Kameen to distinguish between ‘good’ and ‘poor’ writers, where good writers produced a larger amount of passive constructions. These three indices were also selected in Kyle’s (2016) research.

Among other less-frequently-used but more-specific measures that are recommended by Wolfe Quintero et al. (1998) one can mention the IndC/T index or the number of independent clauses per T-unit, adverbial clauses per clause and per T-unit (AdvC/C and AdvC/T), adjective clauses per clause and per T-unit (AdjC/C and AdjC/T), and nominal clauses per clause and per T-unit (NomC/C and NomC/T). Other specific and fine-grained syntactic measures recommended by them include infinitive phrases per clause and per T-unit (InfVP/C and InfVP/T), participial verb phrases per clause and per T-unit (PartVP/C and PartVP/T), and gerund phrases per clause and per T-unit (GerVP/C and GerVP/T). Two further developmental indices in SLA were also proposed by them as definite articles per clause and per T-unit (DefArt/C and DefArt/T) and their counterpart, indefinite articles per clause and per T-unit (IndefArt/C and IndefArt/T).

Another noteworthy index which was proposed and studied in Loban (1976) is the Index of Elaboration, which is a weighted index of syntax elaboration to analyse “the ways by which the basic subject and predicate are expanded” (p. 18) with features such as adverb, clauses, phrases, appositives, etc. A weight is assigned to each elaborated structure based on the list of weights in his work. In another method, this elaboration is assessed based on the number of grammatical transformations involved in producing a sentence. This latter method is also referred to as ‘syntactic density’ and discussed in detail in Loban’s work.

Golub and Frederick (1971) also investigated an index of ‘single-base transforms’ which they define as “sentences appearing in the form of questions or imperatives, the passive or emphatic voice, expletive, or negative” (p. 13). Other indices investigated by Golub and Frederick include ‘adjectives per noun’ as a ratio of all adjectives to all nouns in a sample, ‘adverbs before the verb’ and ‘adverbs after the verb’, ‘adverbs per T-unit’ as a ratio of all adverbs to all T-units in a sample, and the frequency of ‘adverbs in noun phrases’ considering all types of adverbs , among other indices.

There are also a number of other syntactic measures and analysis approaches that are mainly geared for first language acquisition and development, readability formulas, and the studies on transformational grammar theory. Noteworthiest of them are the Index of Productive Syntax (IPSyn, Scarborough, 1990), the Developmental Sentence Score (DSS, Lee, 1974), Developmental Level Scale (D-level, Rosenberg and Abbeduto, 1987), the Derivational Theory of Complexity measure (Fodor, Bever, & Garrett, 1974), the measure of total and maximal depth (Yngve, 1960), and the Directional Complexity or Dcomplexity (also called ‘Syntactic Complexity Formula’, Botel and Granowsky, 1972). Since these indices are not relevant to the present research, I refer the interested reader to these citations for their extended discussions.

2.3.6. Final Remarks on the Selection and Effectiveness of Syntactic Complexity Measures

Most of the syntactic indices that were discussed have been used in first and second language acquisition and development work e.g., the study of these syntactic measures as developmental indices in Kyle (2016), Ortega (2000), and Wolfe-Quintero et al. (1998). There is a body of research on the effectiveness of various syntactic complexity measures as reliable indicators and/or predictors of syntactic proficiency and as reliable discriminators of proficiency differences in these contexts. However, the studies which employed these indices in proficiency-related research vary significantly in their scope, sample size, mode of language, learner and non-academic vs. academic writing, the number of groups and their English language backgrounds, and whether proficiency was defined by holistic rating or by programme level, etc (e.g., Ai & Lu, 2013; Kim, 2014; Lu, 2010, Lu & Ai, 2015; Ortega, 2003). On the other hand, we have studies that questioned certain measures for specific purposes and studies with contradictory findings in this regard. For example, length-based measures are assumed to relate to fluency and productivity rather than complexity of learners' production in the Wolfe-Quintero et al.'s (1998) study of second language development, while studies like Ai and Lu (2013), Lu (2011), Ortega (2003), and Park (2012) list length-based measures of MLT, MLC, and MLS as significant predictors of proficiency level (as gauged by holistic ratings and/or school and programme levels). There are also discussions of the relationship between proficient L2 writers and long texts based on syntactic indices in Frase et al. (1999) and Grant and Ginther (2000). Inconsistent use of definition and quantification of certain syntactic terms and measures are other instances of issues that complicate the interpretation of the findings of different studies (e.g., see the discussions in Ortega, 2003; Wolfe-Quintero et al., 1998). Furthermore, only a few studies have used these syntactic complexity measures (usually a few of these measures only) in specialised academic writing corpora, such as discipline-specific and genre-specific corpora.

In the presence of such inconsistencies and research gaps, it seems plausible, therefore, to systematically test a large set of syntactic complexity measures, especially those with contradictory findings, using various independent variables such as English language background, task types, the effect of genre and sub-genres, and possibly the effect of learners' L1s to find a consistent pattern which can guide future studies on the selection of the most reliable and relevant indices, especially for academic writing research. This practice that is adopted in this thesis, is in line with the findings and conclusions of many previous studies which have indicated that different syntactic complexity measures and structures reveal different information about the linguistic complexity, proficiency, and development of the students and that different traits and constructs of syntactic complexity may affect the syntactic development towards native-like proficiency in different ways (see for instance the discussions in Cheung & Kemper, 1992; Halleck, 1995; Homburg, 1984; Kuiken & Vedder, 2008a; Norris & Ortega, 2000, 2009; Ortega, 2000; Wolfe-Quintero et al., 1998, among others). Therefore, they recommended the use of both global and specific indices of syntactic complexity at the phrasal, sentential, and clausal levels as well as the assessment of subordination, coordination and overall length-based complexity. The studies that will be reviewed in the next chapter have investigated various lexical and syntactic complexity that were measures reviewed in this chapter regarding their effectiveness as indicators and predictors of linguistic proficiency in the context of SLA, corpus, and academic writing research, often with regard to one or more of the independent variables mentioned above.

3 Lexical and Syntactic Complexity in SLA, Corpus, and Academic Research Studies

3.1. Overview

In the previous chapter, lexical and syntactic complexity dimensions/constructs and their constituent measures were defined and their operationalisations regarding various measurement criteria and quantification methods were discussed. In this chapter, I focus on the investigation of these constructs and measures in various types of research and synthesise the main findings of previous studies in four broad areas. In section 3.2. I review the trajectories of lexical and syntactic complexification in SLA data, including the developmental trends. The main arguments of these complexification trajectories will be used against the findings of this study to interpret the results in chapter six. Section 3.3 is dedicated to non-academic SLA studies that investigated the main lexical and syntactic complexity constructs and their quantifiable measures and their effectiveness in capturing lexical and syntactic development, proficiency (differences), and in capturing English L1 vs L2 texts. These discussions will be based on the construct validity of these complexity measures, especially concurrent and predictive validity of these complexity measures when it comes to their relationship with proficiency and development. The next section, 3.4., follows the same pattern but in the context of academia. In this section, non-specialised corpus-based SLA studies in academic settings will be synthesised for evidence of reliability and validity of these indices as mentioned above. I will then turn to specialised corpus-based studies (e.g., discipline-specific and genre-specific) in section 3.5 to review the handful of studies that reported the effectiveness of these complexity measures as indicators of proficiency or capturing genre and disciplinary differences and present the main findings about what is generally considered as proficient academic writing regarding various linguistic features. Finally, in 3.6 I provide a brief overview on the effectiveness of EAP programmes, especially ESL academic immersion programmes in developing linguistic proficiency and the necessity for incorporating data from such programmes into comparative linguistic proficiency studies.

3.2. Trajectories of Lexical and Syntactic Complexification and Developmental Trends in Productive SLA Data

Syntactic complexity has been characterised as a range of syntactic structures, as well as the amount of sophistication of such structures (Ortega, 2003; Pallotti, 2015). “The origin of syntactic complexity”, Barker and Pederson (2008, p.1) note, “is not completely clear.” They propose that, to some degree, it could be seen as a result of the evolution of the communication system, while attributing cross-linguistic variation to historical and developmental circumstances. The so-called developmental circumstances as noted above, in Dahl's (2004) thesis are the development of grammatical patterns over millennia whereby any linguistic phenomena including these patterns become 'mature’ by passing through several ‘successive stages’ and hence adds to the complexity of a language. In light of evolutionary annals, quicker decision making and survival needs lead to the development of syntactic complexity whereas “developmental accounts describe how verbs representing separate but frequently-connected events may move through stages of paratactic association (coordination) to syntactic complexity (subordination) to complex verb forms like complements” (Barker & Pederson, 2008, p.2).

Writing courses throughout most of the twentieth century focused on sentence construction grammar which gradually progressed into more complex sentence structures via combining and adding dependent clauses and phrases which were believed to improve writing skills (Beers and Nagy 2009). This stage was followed by an era of dominance of higher-level processes such as organisation and planning in the late twentieth century. This was because sentence quality and sentence complexity began to be perceived as independent, i.e., longer, complex sentences were no longer perceived as the best ways to improve writing quality. Soon after, and with the rise of genre and disciplinary variation research in the late twentieth century and early twenty-first century, syntactic complexity was back into the scene, this time with additional indices that gauge syntactic complexity via various phrase level, clause-level, and T-unit level measures and the findings that indicated that more complex syntactic structures could help the expression of complex ideas and complex relationships between ideas; this led to the increased use of such measures in examining English L1 vs. L2 texts, and L2 proficiency and development (Beers & Nagy, 2009, 2011; Green, 2019; Loban, 1976; Ortega, 2000; STEWART & GROBE, 1979).

The relationships between syntactic complexity, proficiency and development are explained at great length in Ortega (2000). In her extensive review, she noticed that the syntactic structures that are acquired late (e.g., in later stages of linguistic development) are considered more complex in what is known as the ‘cumulative complexity hypothesis’; the presence of such structures, therefore, mark higher syntactic proficiency and maturity (see also Di Domenico, 2017). Along this line and based on the cognitive demands and processing load of certain linguistic structures, Bulté and Housen (2012: 36) argue that syntactic subordination structures are ‘cognitively harder to process than other types of syntactic linking’ and therefore acquired later. This view is also linked with the concept of the inherent complexity of linguistic features, or ‘Structural Complexity’ (Housen & Kuiken, 2009).

Regarding the syntactic complexification of L2 writing development, researchers such as Cooper (1976), Monroe (1975), Ortega (2000), Sharma (1980), and Wolfe-Quintero et al. (1998), and Ortega (2000) argue that the L2 developmental stages move from sentence fragments and clauses to an abundance of coordination, then to an abundance of subordination, and at higher levels of proficiency, manifest elaboration through embeddedness and the amount of phrasal complexity and elaboration, for example by frequent use of nominalisation and non-finite verbal forms, as well as a decrease in the number of T-units and sentences and an increase in the length of clauses. Further evidence is provided in several works to show that coordination is higher in non-native English L2 learners who are less-advanced and subordination is higher in higher proficiency levels of L2 as well as English L1 (e.g., Bardovi-Harlig & Bofman, 1989; Chen, Alexopoulou, & Tsimpli, 2019; Grant & Ginther, 2000; Mancilla, et al. 2015; Monroe, 1975). Other studies supported that a greater amount of phrasal complexity structures, nominalisation, phrasal elaboration, noun phrase modifiers, as well as phraseological complexity measures (e.g., based on academic word collocations) are indicators of proficient L2 and/or academic writing (Biber & Gray 2013, 2016; Bulté and Housen 2014; Gray, 2015; Halliday 2004; Liu and Li 2016; McNamara et al. 2010; Paquot 2019). Ferrari (2012: 283) also argues that according to the ‘Developmental Prediction Hypothesis’, competent L2 learners complexify their texts ‘at clausal level through the use of nominalization, rather than merely increasing the number of subordinate clauses’.

As noticed and stemming from SLA studies, there are three main explanations and implications of syntactic complexification. First is the contribution of syntactic complexity to writing quality and that certain complex structures could help the expression of complex ideas and complex relationships between ideas (e.g., Beers & Nagy, 2009, 2011). The second is the developmental trend as discussed earlier. I address the third explanation to syntactic complexification based on the ‘Functional View’ where many studies have acknowledged that linguistic complexity cannot be understood in isolation, but is to be taken as a function of/elicited based on task (e.g., task type, condition, and complexity), genre, rhetorical features, English language background and L1s of the writers, topic, disciplinary norms, etc (Biber, 2006; Biber and Gray 2013, 2016; Ellis 2009; Gray 2015; Lu 2011; Lu et al. 2020). These studies will be reviewed in the following sections.

Lexical complexity is also defined as the amount and/or proportion of content words and diverse lexical items as well as the sophistication of such items regarding their rarity and infrequent use and the amount of specialised vocabulary (Bulté & Housen, 2012; Jarvis, 2017; Kyle & Crossley, 2016). Despite variations in research designs and objectives as well as the operational definitions of these constructs and their constituent measures, a general trend in the main body of works on lexical complexity attests to the trajectory of lexical complexification of L2 production via more use of content words (lexically dense discourse), the diversification of lexis, and more use of less frequent words or rare or advanced words and phrases (e.g., the discussions in Bulté & Housen, 2014; Friginal, Li, & Weigle, 2014; Malvern et al., 2004; Treffers-Daller, Parslow, & Williams, 2016; Yoon, 2017).

Lexical diversity which is considered as the surface-level manifestation and one of the four behavioural constructs in Bulté and Housen's (2012) model of Lexical complexity constructs, is believed to contribute to the systemic lexical complexity, e.g., the elaboration, size, and range of L2 lexical items. Yoon (2017) discusses previous works' results in conjunction with his findings regarding lexical complexification and suggests that lexical sophistication tends to develop at higher proficiency levels whereas lexical diversity tends to develop at lower L2 proficiency levels. I have already reviewed such studies in chapter two that observed the impact of frequency of words on learning, i.e., the observations that high-frequency words are learned and used at early stages of language acquisition and low-frequency words are produced in higher levels of proficiency (e.g., Kyle & Crossley, 2015; Nation, 1990; Perfetti, 1985; Rayner & Pollastsek, 1994; Vermeer, 2004). These cumulative findings have led to the increased use of frequency-based methods to gauge productive lexical knowledge and lexical development and proficiency differences (e.g., Bardel & Gudmundson, 2012; Crossley & McNamara, 2013; Vermeer, 2000 among many others).

In the following sections, I review a large number of corpus-based SLA studies both outside and in the context of academia regarding the relationships between lexical and syntactic complexification and development, L2 proficiency, and English L1 vs L2 differences. In doing so, I also synthesise the main findings regarding the effectiveness of various measures as quantifiable representatives of lexical and syntactic complexity constructs to set the scene for the measure-selection process in chapter five.

3.3. The Effectiveness of Lexical and Syntactic Complexity Measures as Developmental and Proficiency-level Indices: Corpus-based SLA Studies in Non-academic Contexts

Lexical and syntactic complexity, as discussed in the previous chapter, has attracted many types of research studies which target one or more of its constructs and measures to gauge L2 writing and speaking proficiency and development, to examine the effects of task types and conditions, genre, gender, and L1 background on the measures' values, and to find the relationships between these measures and other linguistic indices (e.g., Ai & Lu, 2013; Bardovi-Harlig & Bofman, 1989; Bulté & Housen, 2012; Housen & Kuiken, 2009; Ishikawa, 1995; Kuiken & Vedder, 2012; Lu, 2011; Lu & Ai, 2015; Ortega, 2003 among many others).

A review of these studies, therefore, suggests one strand of research and findings regarding the use (the type, amount, and distribution) of lexical and syntactic measures and structures in English texts produced by students with higher linguistic proficiency vs. lower-level ones, as well as those in the texts of English L1 vs. L2 students. Another strand of research focuses on the complex vs. simple syntactic structures and probes into the use of subordination, coordination, and phrasal-level complexity as well as lexical density, diversity and sophistication in various writing genres and corpora. A brief synthesis of these findings is as follows.

A seminal and prominent work in investigating lexical and syntactic complexity measures mainly as developmental indices in first and second language studies is Wolfe-Quintero et al. (1998) and the authors' review of studies in the context of the CAF framework. In this book, complexity is one of the components in the three-faceted L2 proficiency paradigm of Complexity, Accuracy, Fluency (CAF) proposed by Skehan (1989). The CAF framework is used to describe the written performance of language learners as well as indicating learners' linguistic development through stages of learning (e.g., Bulté & Housen, 2012; Norris & Ortega, 2009; Pallotti, 2009; Skehan, 2009; Wolfe-Quintero et al., 1998, etc). Much earlier studies such as Arthur (1979) showed how an increase in learners' proficiency corresponds with an increase in CAF values.

Skehan was one of the pioneers of including complexity in studies of second language learning and acknowledged it as one of the "useful measures of second language performance" (2009, p. 510). He elucidates the issue in that successful second language performance requires 'complexity' - here defined as "more advanced language", 'accuracy' - "a concern to avoid error", and 'fluency' - which is identified as "the capacity to produce speech at a normal rate and without interruption" (p. 510). Since we have a limited attentional capacity and working memory, committing attention to one area of performance leads to a drop in performance in other areas of CAF. He then argues that the mentioned Trade-off Hypothesis could explain the reasons behind the positive correlation between fluency and accuracy as opposed to complexity in task-based performance studies, in which tasks that require manipulation of information, results in higher complexity. Contrariwise to Skehan's Trade-off Hypothesis, findings of complexity-accuracy correlation supports Robinson's (2001) Cognition Hypothesis which postulates that "increasing the cognitive demands of tasks", task complexity, would "push learners to greater accuracy and complexity of L2 production" (Robinson & Gilabert, 2007, p. 162).

Along the same line, a growing host of research studies explored the relationship between syntactic complexity, grammatical complexity, and/or CAF measures and task, planning, and performance (Crookes, 1989; Ellis & Yuan, 2004; Farahani & Meraji, 2011; Foster & Skehan, 1999; Ghavamnia, Tavakoli, & Steki, 2013; Ishikawa, 2006; Kuiken & Vedder, 2008a; Rahimpour & Safarie, 2011; Salimi, Dadaspour, & Asadollahfam, 2011; Skehan & Foster, 1997) as well as examining the measures against gender and age variables (Naves, Torras, & Celaya, 2003; Waskita, 2008).

In her review of studies on syntactic complexity measures, Ortega (2000) emphasises the importance of these indices on learners' development of certain linguistic features such as grammar, to understand the role various task types play in L2 writing, to recognise L2 text differences, as well as the effect of experimental interventions on the production of certain syntactic structures. Rafoth and Combs (1983) equally regard syntactic complexity as "one of the most reliably measured aspects of writing ability" (p.165). Regarding the application of such measures, Larsen-Freeman (1978) proposed that these could be used as placement criteria in L2 language development bands, and Wolfe-Quintero et al. (1998) acknowledged that these indices could be employed for pedagogical, acquisition and testing purposes in second language studies.

Wolfe-Quintero et al. (1998) also elucidated that even though language development and proficiency level are not equivalent, "measures of language development ought to be able to distinguish between learners at clearly different levels of proficiency" (ibid., p. 118). Before these indices could be employed as indices of L2 proficiency and development in investigations, however, they underline that their construct validity needs to be evaluated via repeated sampling reliability, concurrent validity, and predictive validity. The former criterion is satisfied by the consistency of the measure with different participants and "a consistent, linear progression of the measure according to externally-determined proficiency levels across different studies" (ibid., p. 117), regardless of the ways the proficiency levels are defined, e.g., based on holistic ratings, test scores, programme or school levels, etc. The last two criteria (concurrent and predictive validity) are generally subsumed in psychometrics under the umbrella term ‘criterion validity’. This is mainly to do with the generalisability, i.e., how well the measures can reflect or predict proficiency at the time of the study (e.g., concurrent validity) and at a future time (predictive validity). Concurrent validity would be consistent evidence of significant or positive high correlations between such complexity measures and proficiency. In measure testing, this type of validity is also used to examine how well a particular measure performs compared to an already-established measure (e.g., in this study’s case, how well a measure is correlated with an already-established index of L2 proficiency and/or development). These evaluation criteria promise a formidable task for L2 researchers: not only the previous works come with different flavours of research designs, sample sizes, proficiency or developmental classification criteria, etc, most studies on English L2 proficiency and development also employ only one or a few of such complexity measures, sometimes using different measurement criteria and/or analysis tools. Despite these inconsistencies and relying on the available information from previous research findings, L2 researchers (e.g., Malvern et al., 2004; Kyle, 2016; Ortega, 2000; Jarvis et al., 2003; Verspoor, Schmid, & Xu, 2012 among many others) have found some patterns to suggest certain lexical and syntactic complexity measures are reasonably good indicators and/or predictors of English L2 writing ability and quality, proficiency levels, and developmental stages. These reliability and validity criteria, as well as the cumulative evidence from the scholarly body of research in this area, will be taken as the main criterion/standard to select the lexical and syntactic complexity measures in this study presented in chapter five, section 5.3.1. A concise synthesis of such works appears as follows.

Regarding the effect of genre and text types, Beers and Nagy’s (2011) multi-faceted study investigated the four writing genres of narrative, descriptive, compare/contrast, and persuasive in a longitudinal study of school English learners using the subordination index of C/T and length-based index of MLC. They found that persuasive essays had more subordinate clauses than other genres, and descriptive texts had longer clauses (measured as the number of words per clause). They concluded that syntactic complexity is highly dependent on genre types. Stewart and Grobe’s (1979) work shows that syntactic indices like words per T-unit and per clause significantly correlate with writing quality of fifth graders across task types in expository texts. Genres, task types, and the type and number of indices, therefore, play important and inter-related roles in determining the quality of writing, as Beers and Nagy (2009) underline; “writing high-quality texts in different genres … involve acquiring productive control over genre-specific structures that are tied to the communicative goals of writing” (p. 192). There are also investigations on the effect of genre on lexical diversity measures. Students in Olinghouse and Wilson’s (2013) study, for instance, demonstrated higher lexical diversity in narrative texts than in informative and persuasive texts, as analysed by holistic ratings. Story and persuasive texts also featured higher lexical diversity than informative texts; however, informative texts contained more content words. Among the three genres of story, persuasive and informative texts, vocabulary diversity (measured as MTLD) turned out to be a strong predictor of story texts.

Apart from genre, the effect of discipline on the production of complexity measures was also investigated. Green’s (2019) study of the cross-disciplinary variation of linguistic features in secondary school textbooks is an instance. The findings discriminate humanities from science subjects regarding noun phrases, auxiliary verbs, academic phraseology, and dependent clauses. The results also show a contrast in subjects of history and physics regarding noun phrase complexity, especially features like the prepositional expansion of noun phrases (e.g., prepositions followed by prepositional phrases).

Lexical complexity measures were also subject to various English L1 and L2 text differences and developmental investigations. Lexical density and diversity, for instance, were used to track the lexical development of 10-year-olds through university (Johansson, 2008). Similarly, Durán et al. (2004) set out to track the lexical diversity development of thirty-two English L1 children across ten different ages using the D measure (as analysed via the vocd software, see McKee et al., 2000) where they found a significant developmental trend. They also showed that the D measure can be used as an indicator of ESL/EFL development of learners aged 18-30. A similar conclusion was drawn by Malvern et al. (2004) who demonstrated that the D measure has been an effective measure of language development and maturity in first and second language writing of both children and adults.

Among proficiency-related studies, Treffers-Daller, Parslow, and Williams (2016) also employed several measures of lexical diversity (TTR, the Index of Guiraud, Vocd-D, HD-D, and MTLD) to discriminate between essays of ESL students in different CEFR proficiency levels. They also showed that the students in higher bands of CEFR produced texts with more diverse vocabulary. Among the indices, MTLD showed to be a good predictor of Pearson test scores, and more importantly, that lemmatisation (e.g., taking the lemma as the unit of analysis) had a significant effect on the lexical diversity scores. Similarly, the two measures of MTLD and Vocd-D were also incorporated into the study of Crossley et al. (2011) to examine their predictability in the variance of the human evaluation of lexical proficiency across three proficiency levels of beginner, intermediate, and advanced for English L2 texts as well as a group of English L1 texts. They concluded that these lexical diversity indices are associated with vocabulary size and depth. Kyle and Crossley (2015) also assessed the validity of 40 lexical sophistication indices based on their relationships with two types of language proficiency scores of holistic lexical proficiency and holistic speaking proficiency where the holistic scores were assigned by trained human raters.

A type of measure validation can also be seen at the intersection of programme-based proficiency levels and holistic ratings in Verspoor, Schmid, and Xu (2012). They assessed English L2 teenage learners' texts across five proficiency levels (A1 to B2 according to the CEFR framework) as evaluated via holistic ratings for a variety of syntactic and lexical indices, including sentence-level structures, dependent clauses (adverbial, nominal, relative, and non-finite clauses), verb phrase structures, lexical sophistication measures (a Customized Lexical Frequency Profile, CLFP index), and the lexical diversity index of Guiraud. Their findings also corroborate previous works in that the number of dependent clauses and the values of Guiraud index were robust measures for discriminating between proficiency levels.

Regarding the reliability of various syntactic complexity measures, both Lu (2010) and Yoon and Polio (2016) confirmed the reliability of the syntactic measures in L2SCA and reported high correlations between these measures and human annotation of essays. Polio and Yoon (2018) also investigated the validity and reliability of these syntactic measures further and reported that the measures in this analyser can reliably diffrentiate between genres of argumentation and narration based on human-annotated essays. Lu (2017, pages 505-506) lists a number of studies that reported that the measures in L2SCA are predictive of holistic measures of writing quality. 11 of the measures in this analyser will be investigated in the present research as will be described in chapter five.

The main studies reviewed in this section testify to the effects of texts' genre, topic, task types, age, discipline, and proficiency levels on the values of lexical and syntactic complexity indices as well as on the overall quality of writing in the context of non-academic first and second language acquisition and development. These collective findings of these studies as well as the findings on specific measures show that, overall, lexical and syntactic complexity indices are good indicators of proficiency, e.g., based on the discussion of reliability and validity of these measures as mentioned earlier.

3.4. Lexical and syntactic Complexity Measures in Non-specialised Corpora in Academic Contexts: Developmental and Proficiency Studies

Just as in non-academic studies, SLA studies in the context of academia based on general and non-specialised corpora (e.g., argumentative essays) also render evidence to the effectiveness of various lexical and syntactic complexity measures as indicators and predictors of proficiency, and as indicators of writing quality. Over the past few decades, a multitude of studies in the academic context has addressed the effect of one or several of lexical and/or syntactic complexity indices on writing quality, linguistic proficiency and development (e.g., Ai & Lu, 2013; Gregori-Signes & Clavel-Arroitia, 2015; Doró, 2008, 2015; Kim, 2014; Lu & Ai, 2015; Yang, Lu, &Weigle, 2015).

Continuing from the discussions of reliability and validity (e.g., concurrent and predictive validity) in 3.3, various proposed measures as quantifiable representatives of lexical and syntactic complexity constructs of density, diversity, sophistication, length of production, subordination, coordination, and phrasal complexity have shown to be effective as indicators and predictors of proficiency and in capturing differences in English L1 vs L2 writing across genres and text types in the context of academia as well. The concurrent validity and repeated sampling reliability of some of these measures have been also investigated via a research synthesis in Ortega (2003). Taking sample sizes into account for the measures that showed between-proficiency differences across studies, she addressed the issue of "how different is different enough in terms of magnitudes expressed in readily interpretable units" (p. 498). The effectiveness of these complexity measures is shown in the following sample studies.

Syntactic complexity indices have been investigated at great length and depth in English L2 writing in the academic contexts (e.g., Ai & Lu, 2013; Lu & Ai, 2015; Ortega, 2003; Vyatkina, 2013; Yang, Lu, & Weigle, 2015 among many others). Some studies have investigated syntactic measures in the ESL academic contexts (e.g., Flahive & Snow, 1980; Larsen-Freeman, 1978, 1983; Bardovi-Harlig & Bofman, 1989; Homburg, 1984; Perkins, 1980; Yang, Lu, & Weigle, 2015 among others) while others were carried out in EFL academic contexts (e.g., Hirano, 1991; Nihalani, 1981; Yoon, 2017, etc). These studies, as mentioned earlier, vary with regard to the corpus size, texts' length, genre-related characteristics such as topic, sub-genre types (e.g., narrative, argumentative, etc), and whether the samples were drawn from naturally-occurring texts or were written under examination conditions (e.g., the Michigan Test of English Language Proficiency in Homburg, 1984) or writing placement tests (e.g., Bardovi-Harlig & Bofman, 1989). Despite these variabilities, their results point to the fact that generally, the ESL groups surpassed the EFL groups regarding the values of several syntactic complexity measures (MLS, MLT, MLC, T/S, C/T, and DC/C) which could be attributed to, as Ortega (2000) points out, the higher initial proficiency levels/benchmarks for the ESL groups as requirements to enrol at English L1-speaking universities, for instance, or attributed to the role of input (e.g., quantity and quality of input).

Concerning proficiency as a variable in such studies, Ortega (2003) considers "syntactic complexity measures as indices of college-level L2 writers' overall proficiency" (p. 492) while Wolfe-Quintero et al. (1998) concluded that the effect of proficiency on syntactic complexity values is noticeable when the proficiency is defined as programme level and less significant when it is defined based on holistic ratings. A construct-based synthesis of the effectiveness of such measures is as follows.

The syntactic construct of 'length of production unit' as labelled with measures of MLT, MLC, and MLS in L2SCA (L2 Syntactic Complexity Analyzer, Lu 2010) was shown to be effective in capturing English L1 vs. L2 writing differences of university students (Ai & Lu, 2013; Lu & Ai, 2015). Ai and Lu (2013), for instance, showed that English L2 students produced shorter clauses, T-units, and sentences in argumentative and expository essays compared to English L1s. The same pattern is seen in Lu and Ai (2015) in the combined English L2s with different L1s.

Length-based measures also were shown to be good indicators and predictors of English L2 writing proficiency differences (Ai & Lu, 2013; Kim, 2014; Lu, 2011;). Lu (2011) for example showed that MLC and MLT values linearly increase across three EFL proficiency levels in their argumentative essays. Kim (2014) also found that MLT is a strong predictor of English L2 writing proficiency. In developmental studies, Ortega (2003) observed, that 'mean length of T-unit' changes substantially in both EFL and ESL texts. In proficiency-related studies, she concluded that MLC and MLT indices were reliable indicators of L2 writing proficiency differences. MLT also showed a strong positive correlation with the writing quality of ESL students as scored by human raters in Yang, Lu, and Weigle (2015).

These length-based measures were further investigated by Yang, Lu, and Weigle (2015) where they were shown to significantly reflect the scores by human raters regarding the quality of argumentative essays.

Subordination measures were found to have good distinguishing power for English L1 vs. L2 writing (Ai & Lu, 2013; Lu & Ai, 2015). Ai & Lu's (2013) comparative corpus-based study showed that English L2 students produced relatively smaller amounts/proportions of subordination structures compared to English L1s. The same results are obtained from English L2 groups with different L1s in Lu and Ai (2015) regarding the CT/T, DC/C, and DC/T measures.

With regard to subordination indices as indicators and predictors of proficiency, the results of several studies indicate differences between the English L2 writings of lower and higher proficiency levels (e.g., Ai & Lu, 2013; Grant & Ginther, 2000; Kim, 2014; Ortega, 2003). CT/T (complex T-units per T-unit), for instance, was found to be a strong predictor of English L2 writing proficiency in Kim (2014); both DC/C and DC/T also showed a linear increase across English L2 proficiency levels in Ai and Lu (2013). In Yoon's (2017) corpus of college-level argumentative essays, however, clausal-level changes across proficiency levels were marginal (e.g., corroborating the findings of Lu, 2011 and Bulté & Housen, 2014). Ortega's (2003) research synthesis also shows that the C/T index is a reliable indicator of proficiency-level differences of L2 writing. A greater amount of subordination is also linked with higher-rated L2 writing based on human ratings (e.g., Grant & Ginther, 2000).

The indices that quantify coordination have received mixed results. In Ai & Lu (2013) English L2 university students produced similar amounts of coordination (e.g., sentential coordination) to English L1 students, but differed in the amount of phrasal coordination as measured via CP/T. The lower proficiency EFL students in their study also produced more coordination per clause (CP/C). This is in contradiction to the results of Lu and Ai (2015), where the values of both of these measures were larger in the English L1 group's argumentative essays.

With respect to coordination as a distinguisher of proficiency-level differences, Lu's (2011) study of college-level argumentative essays of Chinese EFL learners showed that the values of both CP/C and CP/T indices linearly increases across three proficiency levels.

Various measures of phrasal complexity were also employed in such SLA studies in the context of academia where they were shown to discriminate between English L1 and English L2 texts (Ai & Lu, 2013; Lu & Ai, 2015) and to discriminate between proficiency levels (Lu, 2011; Kim, 2014). Complex nominals of CN/T and CN/C, for example, were higher in English L1 essays in Ai and Lu (2013), Lu and Ai (2015). The proportion of verb phrases as measured via VP/T was higher in English L2s in Lu and Ai (2015). In Kim (2014), however, the same measure shows significant differences across proficiency levels. These results will be revisited in the discussion of the findings of the present study in chapter six.

The values of the same measures of complex nominals also linearly increased from low to high-proficiency levels in Ai & Lu (2013) and Lu (2011). CN/C, for instance, is found to be a reliable indicator of L2 proficiency and development at non-adjacent proficiency levels in Yoon (2017). CN/T was also confirmed to be a strong predictor of English L2 writing proficiency in Kim (2014); in her study, both CN/T and CN/C values significantly differed across the proficiency levels. Phrasal-level measures were shown to be reliable indicators of writing proficiency in Yoon (2017). Lastly, higher-scored essay samples (scored by human raters) in Yang, Lu, and Weigle (2015) contained greater amounts of complex noun phrases.

Reaching a holistic picture of syntactic complexity in L2 writing is a formidable task due to the multiplicity of approaches, research designs, sample sizes, and quantification methods used in various studies. However, synthesising the discussed research studies on syntactic complexity in this section and the previous section provides some consistent patterns which substantiate the claims that coordination structures are used in earlier stages of English learning and subordination structures are used in intermediate to advanced stages and hence the values of their representative indices (e.g., Ai & Lu, 2013; Bardovi-Harlig & Bofman, 1989; Cooper, 1976; Crossley & McNamara, 2014; Mancilla, Polat, and Akcay, 2015; Monroe, 1975; Norris & Ortega, 2009; Ortega, 2000; Sharma, 1980; Wolfe-Quintero et al., 1998). These findings seem to be consistent across SLA, learner English proficiency and development, and academic studies and the comparison between English L1 and L2 production as indicated in these cited works. Bulté and Housen (2012) also conclude that most syntactic structures/measures based on subordination could be considered as ‘hybrid’ measures in that they capture both syntactic diversity and depth, as well as syntactic ‘difficulty’. As mentioned earlier, they concluded that syntactic subordination structures are “cognitively harder to process than other types of syntactic linking” (p. 36). However, syntactic subordination structures have limited applicability as measures gauging linguistic/syntactic development as they only gauge sentential-level complexity (e.g, embedding through subordination) and not clausal and phrasal levels. Therefore, clausal and phrasal-level measures also need to be incorporated into research studies on L2 writing. Syntactic complexification of English L1 vs. L2 academic writing proficiency and development (e.g., Bardovi-Harlig and Bofman 1989; Ortega 2003; Ai and Lu 2013; Lu and Ai 2015 among many others), as well as syntactic complexity differences in English L2 academic writings (e.g., Lu 2011; Kim 2014; Yoon 2017), are testaments to the increased phrasal complexity not only in the academic writings of English L2 to English L1, but also from lower English L2 proficiency levels to higher levels. Length-based measures of syntactic complexity have been also suggested as reliable indicators and predictors of proficiency (differences) of college-level L2 writing in the studies that were synthesised by Ortega (2003).

Similar validation evidence can also be found in studies that investigated lexical complexity constructs and their quantifiable measures.

Lexical density has been investigated in several proficiency-related and development SLA studies in the context of academia (mainly undergraduate writing) using non-specialised corpora (Doró, 2008; Gregori-Signes and Clavel-Arroitia, 2015; Kim, 2014; Šišková, 2012; Vaezi and Kafshgar, 2012). Doró (2008), for instance, found a significant correlation between lexical density and the productive vocabulary test scores of third-year EFL undergraduate essays. She also found a difference in the values of lexical density between argumentative and expository genres of essays. Lexical density is also found to be a strong predictor of L2 writing proficiency and a good discriminator of the three proficiency levels (a linear increase across levels) in Kim’s (2014) study of EFL university students’ essays.

The effectiveness of the measures that represent the construct of lexical diversity was investigated in corpus-based SLA studies in academic contexts as well (e.g., Gonzalez, 2013; Kim, 2014; Šišková, 2012 among others). Lexical diversity measures of MTLD and vocd-D were, for instance, used to analyse 104 ESL and 68 NS university students’ academic writing (Gonzalez, 2013) where lexical diversity showed a significant effect on writing scores, and NS’s lexical proficiency was found to be significantly higher than the ESL group. MTLD was also shown in McNamara et al. (2010) to be a strong predictor of group membership and differentiator of low vs. high proficiency English L1 academic texts. The NDW (number of different words) index is another measure that is reported in Kim (2014) to be a strong predictor of L2 writing proficiency.

Finally, lexical sophistication indices were subject to different types of validation studies as indicators and predictors of proficiency (differences) and discriminators of English L1 vs. L2 texts. As elaborated in chapter two, two main types of sophistication indices, based on externally defined bands and based on less-frequently-used words as filtered against word lists, have been examined in corpus-based SLA studies in academic contexts. Word frequency (based on CELEX), for instance, was used in Gonzalez (2013) and McNamara et al. (2010). It was found to be a strong predictor of proficiency levels in McNamara et al. (2010). In the same year, Lexical Complexity Analyser (henceforth LCA; Ai & Lu, 2010) was developed which paved the way for the computation of additional ratio-based sophistication measures as were described in chapter two. Lu’s (2012) study, though not based on writing proficiency, validated a large number of lexical complexity measures based on raters’ judgments. Among them are one lexical sophistication and two verb-based sophistication measures (labelled as LS2, CVS1, and VS2 that will be described in detail in chapter five of this thesis) which showed strong correlations with test takers' rankings. Most of the lexical variation measures also showed significant relationships with test takers' rankings. These measures will also be described in detail in the measure-selection process in chapter five. A verb sophistication measure (labelled as VS1 in LCA) was also shown to discriminate well between three proficiency levels in Kim (2014) with a linear increase across the levels.

Both Lu's (2012) study of transcribed oral narratives and Šišková (2012) corroborate the construct-distinctiveness of the three lexical complexity constructs of density, diversity and sophistication which is in line with the theoretical and conceptual understanding of these measures that I elaborated in chapter two.

These scenarios and the in-depth discussions in chapter two corroborate the claims on the effect of the number and type of indices as well as the effect of operational definitions on the relationship between various lexical complexity measures and their effectiveness in distinguishing proficiency levels and capturing group differences in academic texts. This issue persists more in studies using lexical measures than syntactic ones, as there is a relatively high consensus on the operational definitions of syntactic complexity measures in the literature. This brings the discussion back to the salient point made in McCarthy and Jarvis (2010) that lexical diversity "can be assessed in many ways and that each approach may be informative as to the construct under investigation" (p. 391). It is plausible, therefore, to extend this argument to other constructs as well and to employ multiple related measures that have been shown as indicators/predictors of proficiency (differences) in advanced L2 and academic studies. In chapter five, I make this case for including an extended set of lexical complexity measures to examine their effectiveness in capturing between-group proficiency differences in sub-sections (six rhetorical sections) of MA dissertations written by English L1 vs. English L2 (both EFL and ESL) students.

3.5. Lexical and Syntactic Complexity in Specialised Academic Corpora

Compared to the host of SLA studies that have analysed general English writing corpora (e.g., essays, assignments, etc), there are only a handful of works that investigated various linguistic complexity indices in specialised academic writing corpora, including discipline-specific and genre-specific (or sub-genres of) texts. This is particularly an underinvestigated area in terms of 1) the description of various specialised academic writing genres or rhetorical sections, 2) measure-validation, i.e., studies on the relationships between various complexity measures and proficiency and development, and 3) understanding differences of English L1 vs. L2 texts, i.e., the measures that well discriminate between English L1 and L2 specialised texts, particularly when both EFL and ESL learners with various L1s are taken into account.

Pietilä (2015), to my knowledge, is the only study that used complexity measures to analyse a corpus of MA dissertations. She analysed the conclusion sections written by English L1 vs. L2 groups in linguistics vs. literature disciplines using lexical density (the proportion of content words to all tokens), diversity (the type-token ratio, and the D measure), and sophistication (using the LFP and the lambda value in P_Lex software [Meara & Miralpeix cited in Pietilä, 2015]). She found a significant difference between the English L2 groups and the English L1 group regarding lexically sophisticated texts and the proportion of infrequent words which was greater in the English L1 texts. However lexical density and diversity values did not show any such difference. The texts from the two disciplines, however, only differed regarding the proportion of academic vocabulary: the linguistics texts contained a larger proportion of academic-specific vocabulary. Since Pietilä only analysed the conclusion sections of the dissertations, we do not have any evidence to know if these measures would not have shown significant differences among the groups in other rhetorical sections, e.g., abstracts, method, literature review, etc. Furthermore, only a few measures have been used, among which there is the highly-criticised TTR for text-length dependency considering the significant disparity between words (text length) of English L1 and L2 texts in her study. The total words for the English L1s were five times less than English L2s as Finish L1s and more than nine times less for the English L2s as Czech. This flawed methodology, alone, could be an important/main reason for the obtained insignificant results regarding lexical density and diversity differences of these groups' texts that depend on text length. These research gaps and inconsistencies will, therefore, be addressed in the present thesis by incorporating six main rhetorical sections of MA dissertations, equal-length texts, and a variety of lexical complexity measures as will be explained in more detail in chapter five.

Among the few studies that analysed specialised academic writing corpora, Paquot (2019) included several lexical, syntactic, and phraseological indices of complexity in her study for examining the academic writing (research papers on modern languages) complexity differences of three EFL groups that were assigned to any of the B1, C1, and C2 proficiency levels based on the CEFR framework. She found that the values of the syntactic indices of MLC (mean length of clause) and CN/C (complex nominals per clause) and lexical indices of rttr and cvv1 increase across proficiency levels. Additionally, the lv, vv2, and adjv indices' values are found to increase in non-adjacent proficiency levels. These patterns of differences, however, were not statistically significant. These lexical indices all capture lexical diversity and the syntactic indices in her study capture sub-clausal complexity. Her results also show that the highest proficiency group (C2) produced more sohphsiticated texts: all lexical sophistication indices of Ls1, ls2, vs1, cvs1, and vs2 showed larger values for the C2 group. These findings will be revisited in chapter six for interpreting this study's results.

Lu et al. (2020) is a recent attempt at systematically investigating the syntactic features of different rhetorical functions (based on rhetorical moves and steps) in a large-scale corpus of Introduction sections of published research articles in social sciences (including applied linguistics articles). Their study shows significant variation in the use of syntactic complexity indices across rhetorical functions by expert writers. These measures included global measures like sentence length, as well as indices capturing finite subordination, clausal elaboration, and phrasal complexity. This is a promising step in identifying linguistic realisations of various rhetorical sections and sub-genres of specialised academic writing. They list a few studies that adopted this analytical approach using lexical bundles and expressions and emphasised that outside this restricted circle, no study has investigated the relationship between linguistic complexity measures and genre features. This line of research, as also aimed in the present research, further our understanding of disciplinary genre-based writing, and as Lu et al. (2020) suggest, a 'form-function' understanding that can lead to improvements in EAP writing pedagogy. Both Flowerdew (2017) and Lu et al. (2020) also call for corpus-based studies for the linguistic description of specialised and discipline-specific academic writing that are vital for syllabus designers and materials developers in English L2 academic contexts.

On the subject of syntactic complexification and academic writing genres, Biber and Gray (2013), for example, documented how nominalisation has become a unique feature of modern scientific writing, especially academic writing in education, psychology, and history. Similar developmental trends have been discussed in Biber, Gray, and Ponpoon (2011) and the dominance of phrasal complexity, especially complex noun phrases in academic writing (research articles). Much earlier, Biber (2006) investigated grammatical variations in academic registers and argued that, overall, dependent clauses are more descriptive of spoken registers than written ones, but passive verb phrases are distinctly descriptive of written academic registers. Disciplinary variation in clausal vs. phrasal complexity was also investigated in Gray (2015) where a trajectory of increased phrasal complexity and decreased clausal elaboration was noticed from humanities to social sciences to hard sciences.

Regarding the characteristics of proficient academic writing expected at higher levels of proficiency, the related scholarly body of work shows that a greater amount of nominal complexity structures, nominalisation, phrasal elaboration, noun phrase modifiers, as well as phraseological complexity measures (e.g., based on academic word collocations) are indicators of proficient L2 and/or academic writing (Banks, 2008; Biber & Gray, 2010, 2013, 2016; Bulté & Housen, 2014; Friginal, Li, & Weigle, 2014; Halliday, 2004; Paquot, 2019). The literature distinguishes between the two written styles in this regard; the dynamic style (e.g., in less formal and oral contexts) and the synoptic style (e.g., in highly-formal academic and specialised written texts) whereby the latter style is characterised by higher lexical density, a greater amount of nominalisation, and longer noun phrases (see for instance the discussions in Biber & Gray, 2013 and 2016 and Bulté & Housen, 2014). Nominalisation, therefore, is not an exclusive feature in phrasal level structures. However, when it comes to the linguistic features of specific rhetorical structures, Lu et al. (2020) show that finite and non-finite dependent clauses were produced significantly more than nominalisation in certain moves such as announcing and discussing the results, presenting research questions, advancing new claims, providing justification, etc. It seems, therefore, to exist insufficient evidence of the dominance of either type of structures (subordination/amount of clausal embeddings vs. phrasal complexity) in higher levels of linguistic proficiency and that these linguistic characteristics vary based the rhetorical functions and disciplines. To recapitulate these and related studies on syntactic and grammatical complexity, academic writing is characterised as structurally more elaborated than speech, contains longer sentences and T-units, features a greater amount of subordinate structures, nominalisations, and phrasal complexity and sophistication, is more explicit (e.g., all logical relations are explicitly encoded in the texts), is more dense and compressed than other types of writing, and is more nominal than verbal (a contrast with spoken discourse). Lexical complexity indices in specialised academic texts are yet to show a consistent result (e.g., Pietilä, 2015 vs. Paquot, 2019).

The characterisation of specialised academic writing texts at higher levels of proficiency, e.g., discipline-specific and genre-specific texts, especially regarding lexical and syntactic complexity measures is limited to a handful of works that were cited earlier. No study so far has also examined such characteristics in postgraduate specialised academic texts based on rhetorical sections and based on various English language backgrounds of the students and the academic contexts, e.g., English L1, EFL, and ESL. This study is, therefore, designed to bridge this gap and obtain a more expansive picture of various lexical and syntactic features of specialised academic texts for the effect of a text-intrinsic characteristic (rhetorical sections as sub-genres of specialised academic texts) and a text-extrinsic characteristic (English language backgrounds of the students based on the academic context) on lexically and syntactically complex texts.

3.6. The Effectiveness of EAP Academic Immersion Programmes on Lexical and Syntactic Complexity of ESL Texts

In chapter one, section 1.4.2, I have already elaborated on the differences between the EFL and ESL academic settings and the necessity for incorporating data from both settings into comparative studies with English L1s. An important work in this area is Ortega (2003). She conducted a research synthesis of L2 writing proficiency across 21 studies that included EFL or ESL academic settings and concluded that ESL writings have been syntactically more complex than EFL texts. She also observed the slower pace of L2 competence in EFL settings which leads to different complexity features in the L2 writings of students in these two contexts.

Compared to the host of works on cross-sectional analyses of lexical and syntactic complexity of academic writing, a relatively smaller number of research studies have probed into the effect of ESL or EAP academic immersion programmes on the acquisition and development of certain complexity indices and subsequently on the lexical and syntactic proficiency (differences) of students. These academic programmes that range from short, intensive ones for specific purposes (e.g., dissertation writing) to long term immersion programmes are usually designed to transition EFL students learning in a non-English context to ESL students that benefit from an authentic and immersive experience in English-speaking countries, oftentimes using the same academic materials as their English L1 peers. Hinkel (2004) reviews several studies on ESL and EAP writing programmes and emphasises that in both undergraduate and graduate academic writing programmes, the knowledge of syntactic structures and vocabulary has been always a top priority and the most-demanded writing skills for English L2 students. She insists that large-scale corpus studies need to be carried out to identify the most-frequent lexical and syntactic patterns of various academic writing genres to help researchers “explain how written academic prose is constructed” and to “inform writing instruction and pedagogy” (p. 52). She demonstrates the discrepancy between what is taught in English for academic purposes programmes and the disciplinary academic writing norms expected of students. This issue, she argues, is rooted in the EAP professors’ unawareness of the “complexities of ESL instruction or L2 learning and acquisition” in the first place. This is while simple tweaks of lexical, syntactic and discourse-level features result in significant improvements in ESL academic writing quality. She maintains that the most influential features in this regard are verb tenses, subordinate clauses and passive constructions; the errors associated with these features are found to obscure meaning and result in lower assignment grades.

The academic-specific features of writing are further discussed in Biber's (2006) renowned book 'University Language' and his characterisation of university registers and the importance of collocations, the expression of stance, lexico-grammatical, and syntactic features in various academic writing genres. Hyland (2016) however, argues that this conformity to academic discourse norms and a rigid focus on conventions may decontextualise pedagogy and lead to "unimaginative and formulaic essays" if "teachers fail to acknowledge genre variation" and "the unpredictable new forms of communication" that are expected from students in their academic careers (p. 18). This constitutes one part of debates among EAP/ESP scholars on whether such courses should focus on disciplinary-specific or register-level features (see for instance the detailed discussions in various works of Hyland and Biber).

Despite the mounting evidence on the necessity of research on specific linguistic complexity features of academic writing in ESL or EAP/ESP programmes, systematic investigations are few and far between. A prominent investigation in this area is Mazgutova and Kormos's (2015) study of an academic writing immersion programme for ESL students in the UK. They reported that lower proficiency ESL students significantly improved in lexical sophistication and all indices of lexical diversity; both low and high proficiency level students also improved in the production of verb variation structures.

With respect to studies that examined the relationship between complexity indices and holistic ratings and human raters, Bulté and Housen (2014) selected a large number of lexical and syntactic indices to compare the values of these measures with subjective ratings of students' overall academic writing quality in an ESL/EAP academic writing programme. This was an attempt to investigate the linguistic indicators of writing proficiency of ESL students during one semester in an academic language programme (an intensive EAP course) using several syntactic measures (e.g., MLS, MLT, MLC compound and complex sentence ratios, coordinate clause ratio, and phrasal complexity) as well as the lexical indices of vocd and Guiraud in a corpus of learner essays. Most of these syntactic complexity indices showed significant increases in their values; by the end of this EAP course, learners produced longer, more complex phrases as well as longer clauses. They concluded that lexical and syntactic complexity "constitute separate, independent dimensions of L2 performance and L2 proficiency, rather than being different aspects of the same L2 performance-proficiency area" (p. 53), supporting the claims of previous scholars such as Skehan (2009a) and Foster and Tavakoli (2009). Crossley and McNamara's (2014) is also among such scant studies that examined the use and pattern of various syntactic structures in a corpus of essays in ESL and EAP academic programmes. They showed the effect of ESL syntactic development on human judgement of writing quality. They also observed that this syntactic development manifests in more nouns and phrasal complexity and that human raters judged clausal complexity as higher quality.

However, as noticed, nearly all studies on the effectiveness of ESL and EAP academic programmes analysed general text types, e.g., essays or writing assignments rather than discipline-specific texts which are the types of texts that are actually expected from such learners in academic settings. To my knowledge, no such study so far has examined various rhetorical sections or sub-genres of specialised academic writings of ESL students. The present research, therefore, takes these research gaps and important linguistic features into account for analysing main rhetorical sections of a discipline-specific academic writing corpus, including the data from ESL academic immersion programmes. The importance of these rhetorical sections as main sub-genres of specialised academic texts will be further elaborated in the next chapter to investigate form-function relationships regarding various linguistic features as well as rhetorical and communicative purposes.

4 Rhetorical Sections in Academic Writing

4.1. Overview

In previous chapters, I established the necessity of investigating the linguistic features of various rhetorical sections as sub-genres of specialised academic texts (e.g., examining the form-function relationships) which has important implications for academic writing research as well as genre-based pedagogical writing practices as strongly recommended by previous scholarship (Flowerdew, 2017; Hyland & Shaw, 2016; Lu, 2017; Lu et al., 2020). That is, this study examines the (co-)occurrence of certain complex linguistic structures (e.g., as specified by distinct constructs) and the rhetorical functions of various parts of scientific academic writing as will be specified in detail in the following sections. This chapter, therefore, will be dedicated to a survey of the characteristics of these sub-genres or rhetorical sections regarding the communicative purposes, rhetorical functions, and linguistic realisations of these functions.

The classification of rhetorical sections in theses, dissertations and research articles in the literature are mainly based on the two proposed patterns of the IMRD structure (i.e., Introduction, Method, Results, Discussion) and the ILMRDC structure (Introduction, Literature review, Method, Discussion, Conclusion). Although a general understanding of the IMRD organisational structure for scientific works has existed for millennia (e.g., in the works of Ibn Al Haytham (also called Alhazen), Ptolemy, and more recently Newton; the evidence for this is presented in appendix A), its use in the modern scientific writing is believed to be originated by the works of Louis Pasteur in the latter parts of the 19th century; it finally became standard in 1972 after the publication of ‘the American National Standard for the preparation of scientific papers for written or oral presentation’, a.k.a the ‘ANSI’ standard (see the discussions in Day, 1989). As will be discussed in the following sections and due to the increasing demand for documenting various types of scientific writing genres, other rhetorical functions and organisational patterns were proposed by subsequent researchers.

The specifications of the main rhetorical sections of academic writing, particularly the sections that are traditionally used in theses, dissertations, and journal articles, are mainly based on the rhetorical characteristics, classification of moves, and the organisational patterns of academic writing genres and sub-genres, e.g., in Bunton (1998), Hyland (2004; 2008), Hyland and Shaw (2016), Swales (1990, 2004) and Thompson (1999, 2002, 2012, 2016), among others. These characteristics, moves, and patterns in turn, substantiate the distinct nature of the rhetorical sections that are traditionally classified into abstract, introduction, literature review, methods and methodology, results and discussion, and conclusion sections and are widely used and unanimously adopted as the *de facto* structure in most theses/dissertations, research articles and conference papers. Therefore, in light of the most recent findings in genre and rhetorical analysis studies presented throughout this chapter (e.g., the division of rhetorical sections in theses in Bunton, 1998, pp. 111-115), in the present study I classify the six main sub-sections of MA dissertations as distinct rhetorical sections as reiterated in chapter five, section 5.2.3, and present the relevant information and previous studies for each rhetorical section in this chapter, sections 4.2 to 4.7.

4.2. Abstract

Several notable studies consider the abstract section of a thesis/dissertation or research article as a distinct sub-genre of academic writing which is characterised by a lexically dense outline, and a summary of the whole thesis/article or as Bunton (1998) describes, as a microcosm of the thesis (Bhatia, 1993; Gillaerts & Van de Velde, 2010; Lorés, 2004; Sánchez, 2018), the presence of specific structural markers and a language that is “objective, clear, and formal” (Ramires, 2017, p. 17). Bitchener (2010), Pho (2008), and Weissberg and Buker (1990) further believe that the function of an abstract is to give the objectives of the study, along with brief statements of the content, methodology, findings, and general or specific implications and contributions of the study; in other words, abstracts are expected to reflect the general IMRD structure of the rest of the article/thesis (Lorés, 2004) or the IPMRC structure (Introduction, Purpose, Method, Results, Conclusion; proposed by Hyland, 2000). The importance and functions of abstracts do not end here. They are unanimously considered as one important criterion for communicating the scientific research to readers and to invite them to continue reading the rest of the article. e.g., by persuading the readers that the rest of the work is interesting, relevant, and the results are reliable and significant (see for example the discussions in Bunton, 1998; Gillaerts & Van de Velde, 2010; Hyland, 2002; Safnil, 2014, and Sánchez, 2018 among others). Besides, the quality of abstracts (especially those written in English) is particularly important as they appear in the abstracting and indexing of publishers (Salager-Meyer, 1992; Thyer, 2008 cited in Safnil, 2014) e.g., the indexing of thesis/dissertation abstracts by ProQuest (PQDT A&I). The quality of abstract is also one of the main criteria in accepting/rejecting conference papers and research proposals (Lorés, 2004).

For the mentioned reasons, in recent years a considerable body of research has focused on different aspects/characteristics of abstracts in thesis/dissertation, conference papers, and research articles, e.g., the linguistic, stylistic, communicative, metadiscourse, moves, genre and structural characteristics (see for instance Bhatia, 1993; Bunton, 1998; Gillaerts & Van de Velde, 2010; Golebiowski, 2009; Hu & Cao, 2011; Hyland & Tse, 2005; Jalilifar & Vahid Dastjerdi, 2010; Jiang & Hyland, 2017; Lorés, 2004; Pho, 2008; Salager-Meyer, 1992; Samraj, 2005; Swales, 1990, and Tseng, 2011 among others). Hyland and Tse (2005), for instance, investigated the frequencies as well as forms and functions of evaluative *that* in the research article, MA dissertations and PhD theses abstracts written by English L2 writers to understand how they thematise attitudinal meanings.

Compared to the wealth of research on various metadiscourse and move analysis of abstracts, a relatively-smaller body of research investigated specific linguistic features, especially prominent lexical, grammatical, and syntactic features/characteristics of abstracts (e.g., Allison, et al., 1998; Bunton, 1998, 2005; Egbert & Plonsky, 2015; Pho, 2008; Yoneoka & Ota, 2017). Yoneoka and Ota (2017), for instance, revealed that, despite similarities between low-quality and high-quality abstracts (assessed by two reviewers via a risk-of-bias tool) in terms of the amount of lexical diversity, high-quality abstracts contain shorter sentences, longer words, a small proportion of verb phrases and a larger proportion of noun phrases. Other linguistic features of tense, voice, stance words, nouns, modal and reporting verbs, that-complement clause, first-person pronouns, as well as hedgers and boosters were also investigated in the works of Egbert and Plonsky (2015), Hu and Cao (2011), Muangsamai (2018), Pho (2008), and the works of Salager-Meyer (1992) and Tseng (2011). Bunton (1998, p. 72) who views an abstract as “a self-contained piece of discourse” representing “some of the best writing of the author”, studied PhD theses abstracts for lexico-grammatical accuracy and lexical and syntactic differences, as a continuation of the works of James (1984), Lewkowicz and Cooley (1995), as well as the study of Allison et al., (1998) on lexico-grammatical analysis of postgraduate writing, especially dissertations.

4.3. Introduction

In his seminal work, Swales’ CARS model (Create A Research Space, 1990) for the analysis of the introduction section (of research articles) that involved 3 main moves of ‘establishing a territory’, ‘establishing a niche’, and ‘occupying the niche’ served as a pivotal guideline for move analysis studies for years to come. This work together with the studies of Dudley-Evans (1986), Bunton (1998, 2002), Kwan (2006) and Bitchener (2010), and other notable works in this area suggest that the Introduction section of a dissertation/article can be viewed as a distinct rhetorical section (Bhatia, 1993 calls it a distinct genre in research articles) with unique communicative characteristics which focus on introducing and/or a background to the study along with the aims, significance, and the structure of the (following sections/chapters of the) articles/dissertations. Swales (2004), for instance, believes that the introduction sections in research articles in the twentieth century "have taken on the create-a-research-space character of the CARS model" (p. 216 and 226). This 'space', he maintains, is a unique environment to show off 'originality', as well as a space for situating the author's research amidst 'a big world' and 'big names'.

After abstracts, the introduction sections of research articles and dissertations have attracted more qualitative and empirical studies compared to other rhetorical sections discussed in this chapter. Notable works on the introduction section include the study of Bhatia (1997a) on the function and structure of introductory genres of academic books; Bunton's (1998) project on the genre and rhetorical analysis of PhD thesis introductions; the work of Joseph, Lim, and Nor (2014) on forestry research introductions; the metadiscourse evaluation of identity in EFL and ESL writers' RA introductions in the 2014 study of Rahimivand and Kuhi; the study of Samraj in 2005 on disciplinary variation in academic writing in the fields of conservtion biology and wildlife behavior; the investigation of rhetorical structure of RA introductions in agricultural science in Shi and Wannaruk (2014); Swales and Najjar's (1987) study on RA introductions in the two fields of physics and educational psychology and the amount of variation in rhetorical features across these disciplines; the move analysis study of Kanoksilapatham in 2005 on biochemistry research articles' introduction section; Nwogu's (1997) work on the struture and function of introductions in medical research papers; and West's (1980) investigation of that-nominal constructions in the introduction sections of biological RAs.

Among the few works on dissertation introductions, Dudley-Evans' (1986) is the noteworthiest. He found that, unlike Swales' move two (summarising previous research) in RA introductions, the dissertation introduction summaries are part of a general move including summaries of the parameters of the research; he calls this move 'defining the scope of the topic', instead. Dudley-Evans' (1986) and Hopkins and Dudley-Evans (1988) further specify other distinct moves in the introduction sections of the dissertations in the form of a cyclical pattern with the components as statements outlining the variable, description of the previous related research, and an evaluation of the present research. Bunton's (1998) investigation of PhD theses also revealed the distinct nature of the introduction sections. He found that the introduction of theses contains the second-highest percentage of total references (after the literature review sections), and contained headings related to brief history/background to the field, a general review of theories, the objectives and scope of the study, the organisation/outline of the thesis, definition of terms, and other remarks; he also noticed several additional steps to the three-move classification of Swales' CARS model, with noticeable differences between the theses in science and technology and those in humanities and social sciences. Bitchener (2010) also qualitatively analysed the Introductions of a master's dissertation and examined the use of tense, active vs. passive voice, adjectives, first person pronouns and contrasting conjunctions and phrases. The most recent investigation of form-function relationships of linguistic features and the rhetorical functions in texts is the systematic analysis of introduction sections of research articles in Lu et al. (2020) in which they demonstrated how certain synatctic complexity structures are more or less prevalent in sentences with specific rhetorical functions based on the revised CARS model.

4.4. Literature Review

The literature review section, as Kwan (2006), Bitchener (2010), and Creswell (2014) note, primarily documents the scholarly works that have been conducted on the general and/or specific topic to establish a gap to be covered by the writer. This section also highlights the value of the study and raises the shortcomings of previous works as well as providing a framework for comparing the results with the findings of other relevant studies.

The analysis of rather long literature review sections of various academic writing genres and sub-genres is scarce, even though this section is part and parcel of most academic writing genres, especially theses and dissertations. Bunton (1998) and Stubbs (1994) early on criticised the prevalence of analysis on 'short texts' and stressed the necessity of including long texts in the analyses of rhetorical structures, genre moves, and linguistic patterns of academic writing. Stubbs (1994) for instance, emphasises that "some patterns of repetition and variation are only realized across long texts" (p. 217). Commenting on the restricted format of IMRD, Bunton (1998) also argues that in this format which is mainly relevant to research articles, the literature review section is assumed as part of the introduction section; he then argues that literature review needs a separate section in theses, and hence respective rhetorical analyses. In practice, however, it is more common to see more than one chapter with different headings (other than the term 'literature review') to comprise the review of the related literature in theses and dissertations. In his investigation of PhD theses, Bunton (1998) also found that these names may vary to ‘background’, ‘literature review’, and ‘theoretical framework’ or other headings depending on/reflecting the specialised nature of that dissertation/thesis. He also found obvious deviations in the PhD theses from the so-called standard format of IMRD, in that theses contained separate literature review and conclusion sections. Similarly, Bhatia (1993) believes that the literature review section deserves a separate section/chapter as it reports on a synthesis of previous research and demonstrates an author’s knowledge of the relevant literature.

Bitchener (2010) is a notable instance of analysing the literature review section of a master’s dissertation by identifying the key functions and thematic structures and organisational patterns. He highlighted that even though these structures "vary from thesis to thesis, it will always contain an introduction, a body and a conclusion" (p. 61). Similarly, Kwan (2006) systematically investigated the literature review (LR) sections of doctoral theses written by English L1 students in applied linguistics. She also observed the presence of introduction-body-conclusion structure throughout the literature review sections and found various thematic sections with recursive move structures in the body parts. She also discovered three additional move elements of ‘relevancy-claiming’, ‘strength claiming’ and ‘the synthesizing of the theoretical framework’ to those identified in Bunton’s (2002) CARS model of generic moves in PhD thesis introductions. She concluded that there are noticeable structural differences between the introduction and literature review sections, and hence the need for separate move analyses with specific attention to cross-disciplinary variations.

4.5. Method and Methodology

The methods and methodology section is specific to empirical types of articles/theses (e.g., see Swales, 2004), and as Creswell (2014) Lim (2006), and Bitchener (2010) indicate, deals with the specifics of research methods and design including data collection and analysis and an interpretation scheme/framework for understanding the results. The nomenclature for this section includes ‘the study’, ‘method’, ‘data and methodology’, and ‘setting and methodology’ as well (e.g., Swales, 2004, p. 219). Lim (2006) argues that this section is crucial in persuading the readers about the validity of the means to obtain the study results. Regarding its relative importance, Swales (2004) cites Berkenkotter and Huckin (1995) to demonstrate how the size and importance of methods section of research articles are reduced in the twentieth century, compared to the introduction sections. He further believes that the methods and results sections account for the main disciplinary differences, among other rhetorical sections of research articles and that method sections in dissertations are often more detailed (which he refers to 'slow' or 'elaborated' methods) and contain discursive method discussions, especially in science and engineering. In social sciences, on the other hand, one can detect the use of "purposive, justificatory statements (In order to control for X, we did Y)" (Swales, 2004, p. 114).

Various types of research ranging from the rhetorical structure analysis, genre moves, metadiscourse features, and linguistic (e.g., lexical, grammatical, syntactic) features were also conducted in method sections. Some instances of such studies are Shi and Wannaruk's (2014) analysis of method sections in agricultural research articles, Nwogu's (1997) analysis of method sections in medical research papers, Kanoksilapatham's (2005) move analysis of biochemical research articles, Rafiei and Modirkhamene's (2012) study of thematicity in the method sections of Iranian students' MA theses/dissertations; and the analysis of that-nominal constructions in the method sections of biological research papers in West (1980) among others.

Concerning the linguistic features of the method section, Swales and Feak (1994) investigated the use of imperative verbs, the past passive and active, as well as sentence connectors; Lim (2006), however, noticed the use of temporal adverbials, compositional verbs, procedural verbs (e.g., in collecting data), and verb phrases pre-modified by adverbs (e.g., as the description of sampling techniques). The frequency and pattern of the use of 34 epistemic lexical verbs were also explored in Dontcheva-Navratilova's (2018) analysis of method sections in linguistics and economics RAs; she found evidence of disciplinary variation in the use of judgment and evidential epistemic verbs. Nominalisation, especially that-nominal constructions are among other linguistic features that were also investigated in the method section of biological scientific articles in West (1980); he found that the method sections contain fewer that-nominal constructions compared to other rhetorical sections. Thompson's (2002) study also investigated the use of modal auxiliary verbs in the method sections of PhD theses and revealed that the modal verbs are used less-frequently in the method sections than the results and discussion sections.

4.6. Results and Discussion

The results and discussion section is sometimes split up into separate chapters/sections and sometimes is represented in one large section where the result of each research question is followed by its interpretation and the comparison of other studies' findings (for a detailed discussion see Swales, 2004, pp. 224-226, and the categories of rhetorical sections in

Thompson, 2002 and West, 1980). Where the discussion section is a separate chapter, the findings are usually reported in the form of summaries especially in the beginning parts (Bitchener, 2010; Pojanapunya & Todd, 2011), and the communicative moves extend beyond a simple presentation and discussion of the results and it usually encompasses a review of the aims of the research as well as theoretical and methodological considerations (Basturkmen, 2009). The results are expected to be presented in the form of tables and figures and the findings are discussed to show the trend of the author's reasoning in light of the results of others' works (Woodford, 1976 cited in Swales, 2004); the presence of persuasive moves and explanation of the significance and accuracy of the data are other noticeable instances dominant in results sections (Swales, 2004).

Compared to the host of research on abstract and introduction sections, studies on different linguistic and structural characteristics of the results and discussion section are rather infrequent. The leading works in the twentieth-century include Brett's (1994) study on results sections and comparisons with discussion sections in sociology Ras, Dudley-Evans' (1986) study of discussion sections of M.Sc dissertations, the study of Holmes (1997) on the RA discussion section in sociology, political science and history, the study of results sections of medical research papers in Nwogu study in 1997, as well as the work of Hopkins and Dudley-Evans in 1988 on the description of discussion sections in research articles and dissertations among other earlier studies. Recent analyses of the results and discussion section(s) mainly turn the spotlight on the research articles (RA) in various disciplines. This line of studies includes Shi and Wannaruk's (2014) study of RAs in agricultural science, Amnuai's (2017) move-analysis study on the discussion section of RAs in accounting, Amnuai and Warranuk's (2013) work on rhetorical move structure of RA discussion sections in applied linguistics, Yang and Allison's (2003) genre analysis on the discussion sections of RAs in applied linguistics and the proposed framework for identifying rhetorical moves in this section, the study of Dobakhti (2016) on the generic structure of discussion sections in qualitative vs. quantitative research articles in applied linguistics, the disciplinary variation in communicative moves in the discussion sections of research articles of English L1 vs. L2 writers in Peacock (2002), as well as the move analysis study of Kanoksilapatham in 2005 on biochemistry research articles.

Among the few studies which have been conducted on theses and dissertations, Thompson's (2002) study observed a larger amount of modal auxiliary verbs used in the discussion sections of the PhD theses in agricultural botany and food economics, compared to the method and the results sections. Hopkins and Dudley-Evans' study in 1988 on the discussion section of dissertations also revealed that the choice of moves depends on the satisfactory results or otherwise, obtained by the student (e.g., statement of result vs. unexpected outcome moves). They also observed that the emphasis in this section is more on the interpretation of results and the way the writer relates them to previous studies in the literature. Iranian EFL students also tend to use different move frequencies to non-Iranian EFL students, according to Salmani Nodoushan and Khakbaz (2011) study of move analysis of MA dissertations. This discrepancy, as Swales and Feak (1994) argue, could be due to the nature and type of research questions as well as the headings and sub-sections that the writers prefer to include in this section. Bitchener's (2010) case study is among the very few linguistic investigations; he examined hedging and the use of simple past tense in reporting quantitative results and the use of hedge verbs, adjectives, and modal verbs in the discussion of results.

4.7. Conclusion

Research into the functions, moves, and linguistic features of the conclusion section, however, portrays a number of different functions and strategies which are often envisaged as 'everything else' needed to be mentioned. Bunton (2005) and Bitchener (2010) list a number of functional moves and steps, prominent among them are the restatement of the objectives of the research, a summary of research findings, the significance and necessity of the study, limitations of the study, recommendations and suggestions for future research, and implications and applications. Dudley-Evans (1994) also agrees that the conclusion sections mainly summarise the main claims and findings before moving on to the recommendation for future studies. Among other early works, Peng (1987) refers to the conclusion section in which "deductions and implications of a wider nature are presented" (p.112).

It ought to be mentioned that in some theses and dissertations, discussion and conclusion sections are combined into one chapter, and hence the respective functions and moves (see the discussions in Bunton, 1998, Yang and Allison, 2003, as well as Bitchener 2010). Bunton (2005) argues that in research articles based on the IMRD format, the conclusion section is usually incorporated into the discussion section and presented towards the end of the article. For instance, the three move structures of 'limitations', 'recommendations' and 'final conclusions' that are conventionally part of a separate 'conclusions' section in dissertations/theses, in the research articles are usually subsumed in the 'Discussion' sections (see for instance the structure analysis and moves in Dudley-Evans, 1994 and Swales and Feak, 1994, and the related discussions in Yang and Allison, 2003).

Bunton (1998) however, based on his empirical findings maintains that conclusions need to be separated in theses and hence, the respective rhetorical analyses. This was later on confirmed by Paltridge (2002) in which he specifies a separate conclusion chapter for each of the four thesis types. In his study on PhD theses, Bunton (1998) further noticed that examiners commented negatively about the absence of a separate conclusion section and marked it as a major structural weakness. Offering a possible explanation, Yang and Allison (2003) noticed that one reason for the absence of headings to indicate the 'conclusion' section was that authors preferred other terms such as 'summary and implications', 'summarizing the study', etc; thus only 65% of the headings in their corpus of RAs included the wordings such as 'conclusion', and 'conclusion and pedagogic implications', etc. They pinpoint a difference in the two sections of discussion and conclusion by arguing that the two sections "differ in terms of primary communicative purposes"; they further indicate that the conclusion sections focus "more on highlighting overall results and evaluating the study" and "pointing out possible lines of future research as well as suggesting implications for teaching and learning" (pp. 379-380).

Among the scant attention that has been given to the conclusion section as a separate rhetorical section, Pietilä (2015) is the only study so far that has investigated lexical complexity measures in a corpus of conclusion sections of MA dissertations; the details of this study has been already mentioned in section 3.5. The study of Adel and Ghorbani Moghadam (2015) investigates the disciplinary variation in the rhetorical move structure of conclusion sections in a corpus of research articles. A cross-linguistic analysis of the conclusion section in applied linguistics RAs is also carried out by Moritz, Meurer, and Dellagnelo (2008) in three corpora of Portuguese, English L1, and English L2. A similar study on the disciplinary variation of intensity markers using linguistics features was also conducted by Behnam and Mirzapour (2012) in their analysis of abstracts and conclusions in applied linguistics and electrical engineering RAs. Bitchener (2010), however, is among the very few who examined linguistic features such as the use of modal verbs and subordination in the conclusion chapter of a master's dissertation. Bunton's (2005) study as a follow-up to his earlier work also confirmed the distinct nature of conclusion sections in a corpus of 45 PhD theses across several disciplines. Furthermore, he cites Paltridge's (2002) survey of guide books in which the conclusion chapter is given separate and distinct status as a chapter in all four types of theses that are distinguished by Thompson (1999).

As elaborated at length in this chapter, the dissertations and theses as longer and more detailed academic writing and/or scientific reports are generally expected to follow a more elaborated structure of AILMRC structure (Abstract, Introduction, Literature review, Methods & Methodology, Results & Discussion, Conclusion or Conclusion & Discussion) rather than the conventional IMRD structure as investigated in many research articles. The additional abstract, literature review and conclusion sections (sometimes under different terms or headings), as discussed in 4.2., 4.4 and 4.7, have been the staple of almost all dissertations and theses across various disciplines. In this thesis, I adopt this distinction for classifying the rhetorical sections of master's dissertations as the academic writing corpus under investigation, and reiterate the main rhetorical and genre moves and structures expected in each rhetorical section in the next chapter, section 5.2.3 as the basis for this classification.

5 Research Design, Methodology, and Methods

5.1. Methodology, Research Design and Objectives

This thesis investigates lexical and syntactic complexity of postgraduate academic writing in a specialised corpus of master's dissertations produced by English L1, ESL, and EFL students in various sub-disciplines of applied linguistics in Iranian and UK universities.

As I pointed out in the Introduction chapter, this study has a four-fold purpose. First, this study will examine the linguistic proficiency differences between English L1 vs. L2 postgraduate writers in terms of the production and the patterns of use of lexical and syntactic structures, and to verify the accuracy of the claims of previous research in this regard. The second goal is to compare the proficiency of the two groups of EFL and ESL where English has been taught and practiced only as a foreign language vs. second language in daily communication in academic contexts to see if the ESL group shows a distinctly different pattern of linguistic features in their texts compared to the EFL group. Third, it aims to find out whether various rhetorical sections as sub-genres of dissertations can be distinctly characterised by certain lexical and syntactic features as produced by the participants of the three groups of English L1, EFL, and ESL. Finally, it intends to test two large sets of lexical and syntactic complexity measures to evaluate the most effective measures as indicators and predictors of linguistic proficiency. This is to build statistical models of lexical and syntactic complexity of academic writing at the postgraduate level as well as statistical models to predict rhetorical section and group memberships based on the selected linguistic indices that are specified in 5.3.1. Since the aspects of this design are inter-related, each of the above issues may be discussed in conjunction with others throughout chapter six in the discussion of results. The related research questions for each phase of the study has already been specified in chapter one, and specific answers to these questions will be provided in chapter six.

The impact of corpus linguistics studies on language learning and academic writing is already taking effect. No longer are pedagogical decisions based on mere intuitions; rather, the study of patterns of a given text or corpus provides important information about the genre and nature of the text as well as the characteristics of its writer. This project is informed by the principles and practices in corpus linguistics in the corpus construction, design, and analysis (see for example McEnery, Xiao, & Tono, 2006; Hunston, 2002). I selected already-written academic texts as instances of naturally-occurring data. I also employed the principle of minimal interference with the texts in the corpus which is crucial for an insight into the authentic linguistic profiles of English L1 vs. L2 students. Deleting parts of the texts is limited to the noise elements as specified in 5.2.4. Following the tradition of corpus-based studies, the first phase of this research is primarily focused on the identification of lexical and syntactic patterns in the corpus, based on the quantitative results and is, therefore, not concerned with the underlying reasons for the presence and absence of certain features in the data (i.e., why each group produces certain linguistic features), nor does it dwell on the solutions to bridge the proficiency gap. However, a number of possible reasons and solutions are offered in the final chapter along with some linguistic examples from the texts in chapter six as a guide and starting point for future researchers who wish to conduct an in-depth investigation from purely a pedagogical point of view. Furthermore, the primary basis of the analysis in this work, as with most corpus-based studies, is the frequency counts: calculating the measures is based on the frequency count of the lexical and syntactic production units as explained in sub-sections of 5.3.

I also follow the methods and practices in quantitative linguistics and statistical modelling for the measure-testing processes that will be described in detail in chapter six.

Investigating the linguistic proficiency differences based on the L1 backgrounds of students is not the focus of this study; some reasons for this exclusion along with recommendations for future researchers in this regard are given in the final chapter.

The setting in which the participants studied and submitted their English academic texts was also an important issue in the design of this study. In the Introduction chapter, I explained how this project was motivated by a re-examination of the claims on differences of academic texts produced by English L1 vs. L2 (EFL and ESL) students as well as the possible role of the learning context, especially academic immersion programmes, on the production of such differences. I also cited several works that have identified differences in linguistic proficiency of students in these two settings (Ortega, 2003), investigations of linguistic proficiency in ESL academic contexts (e.g., Bulté & Housen, 2014; Crossley & McNamara, 2014; Masgutova & Kormos, 2015) and the challenges of academic writing in Iranian EFL academic settings (Derakhshan & Karimian Shirejini, 2020). Here I provide additional information including academic writing practices in these two settings.

In this work, an ESL student is defined as one who studied in an English-speaking setting where English is immediately used as the main language of education and communication in society. Almost all ESL students in this study were required to pass the

IELTS English proficiency test with a minimum score of 7. The ESL participants in this project have studied for their MA course and submitted their dissertations in a UK university and both English L1 and ESL students share the same curriculum, materials, syllabi and modules. This setting also involves the participation of both ESL and English L1 students in various discipline-specific academic seminars, workshops, postgraduate sessions, and conferences within the universities where specialised subjects would be frequently discussed by these groups and experts in the field. MA programmes in the UK are either MA by research or based on taught modules; both types of courses require the completion of a dissertation under the supervision of one or two supervisors. MPhil programmes are also available for the students who wish to connect/advance their MA research to a prospective PhD. The role of supervision in the dissertation writing process varies based on the university and the supervision style of individual supervisors. Most supervisors mainly comment on the content of the dissertations. Most universities also offer research training courses to assist with academic skills, including academic writing. Many such courses do not contribute to the final assessment. I reviewed the syllabi of some of these courses and found that they mainly provide information and assistance in academic writing styles (e.g., coherence, organisation, argumentation) and linguistic accuracy via essay writing. I did not find any information regarding any explicit instruction of complex structures that are associated with proficiency. ESL students in this study did not reside in the UK for a long time, often stayed shortly before and/or for the duration of their postgraduate studies (MA and PhD), often as part of academic immersion programmes that are funded by their home countries.

An EFL setting in this study is defined as an environment where English is not the primary language of instruction and communication within the academic setting and outside academia. Outside academic contexts, attending general English courses in various institutes is the main way that students are exposed to formal English registers and receive teachers' feedback. Students also use various web-based applications outside these classes to advance their English proficiency but these methods are not accompanied by teachers' feedback. In academic contexts, the more widespread use of academic English is restricted to the English departments. In this EFL setting, lecturers are often English L2 professionals and the occasional use of L1 language for translation or disambiguation purposes is widely accepted by both students and lecturers. MA students in Iran have to pass the entrance exam for MA programmes, which, apart from the field-specific questions contains an English language proficiency test with 100 questions with a particular focus on vocabulary and grammar. This test puts a penalty for wrong answers and admission is based on a norm-reference testing process. So although this English proficiency test is not exactly the same as the test taken by the ESL students before admission to the UK universities, the two groups are required to demonstrate relatively advanced knowledge of general English.

Almost all MA programmes in Iran is based on taught modules. Supervisors often comment on the content of the dissertations (referred to as 'thesis' in Iran), usually chapter by chapter but advise the students to work on mechanics of writing and linguistic accuracy on their own. Proofreading is usually limited to obvious spelling mistakes or grammatical errors. Students are usually referred to the previous dissertations written in the department by other EFL students for an insight into the required academic writing style. Academic writing is an essential module in MA courses but mainly revolve around critical essay writing practices. The 'seminar' is also an essential module that prepares MA students for dissertation writing; this module covers the structure of a typical dissertation in detail as well as writing styles and researching/accessing digital media. Students have usually started the research aspects of their dissertations while taking the seminar module. I did not find any information regarding the aspects of linguistic complexity that contribute to proficiency in several syllabi of such courses. In this research, EFL students have lived in Iran (mainly Iranian nationals) and have studied their master's course and submitted their dissertations in an Iranian university under a centralised curriculum and share the same modules and syllabi which were specifically designed to cater for the needs of students learning in non-English environments.

This research is a quantitative and cross-sectional study. The dissertations were reviewed a few times individually and manually for data sorting and cleaning purposes. Some linguistic examples as excerpts from the texts will also be provided in chapter six to discuss the textual understanding of these lexical and syntactic complexity constructs in texts that received low vs. high quantitative values for the representative measures of these constructs. Along with these textual examples, I will also briefly discuss how complexity based on the systems view can be interpreted at a local level, i.e., a sample text, using both quantitative values of the measures and the qualitative understanding of the constructs. That is, where multiple components of a system (here various lexical and syntactic structures) are interconnected in such a way that high values of each and all of these components render the whole system more complex and hence a higher linguistic complexity profile of a given discourse. The functional view, on the other hand, can be best explained based on the statistical modelling methods, e.g., mixed-effect modelling in chapter six, that gauge the effect of rhetorical sections and groups of students on the quantitative values of these complexity measures.

To address the above research areas in the study design, I first constructed an academic corpus of master's dissertations written by 210 English L1, EFL, and ESL students (section 5.2.1) ensuring the representativeness of the corpus in terms of various sub-disciplines of Applied Linguistics (section 5.2.2). Then I divided each dissertation into six sub-genres, each as a rhetorical section based on the literature on genre analysis (section 5.2.3) followed by the text cleaning and preparing process (section 5.2.4), coding and formatting the texts and their respective folders (section 5.2.5), truncating the corpus, and eventually obtaining the final word counts (section 5.2.6). The final corpus then was subject to text pre-processing and analysing using several analysis tools and programmes which calculated two sets of lexical and syntactic complexity measures as possible indicators and predictors of linguistic proficiency in L2 and academic writing (sub-sections of 5.3). Lastly, the obtained values of the measures in each rhetorical section for the three groups were tested using several statistical methods (section 5.4).

5.2. The Construction of an Academic Writing Corpus of Master's Dissertations

An important first step in this project was to construct an academic corpus with the already-written texts as opposed to the ones obtained under testing conditions. This natural approach to data collection, as suggested in 1998 by Bunton, helps avoid conscious manipulation of linguistic structures by students if the objectives of the data collection were known. The following sub-sections give detailed information on the corpus construction processes.

5.2.1. Description of the Data, Data Collection Process and Sampling Methods

In this study, master's dissertations have been collected as the data over a period of three years. To reduce the variability in the data due to time-related factors, only the dissertations which were published/submitted within eight years prior to the commencement of the final data analysis in this study were collected. By the same token and to limit the effects of participant variables, the corpus was drawn from female and male participants of the 23-35 age group. Both quantitative and qualitative original and empirical research studies were chosen whose sections were the same as/similar to the scheme in 5.2.4. Since master's dissertations were not freely available in large numbers, several data collection and sampling methods were used:

- First, cluster sampling was used to identify and sample from major universities/departments where TEFL and Applied Linguistics courses are offered to master’s students in Iran and the UK.
- Second, random sampling was used to obtain random samples of dissertations from each identified cluster; dissertations were collected in this method from universities’ authorities in Iran who had permission to distribute dissertations for research purposes. These dissertation writers have already given permission for their works to be used for research purposes. Additional dissertations were shared by individual students.
- Third, data collection messages along with a Participant Consent Form were distributed on various websites and via several universities’ mailing lists in the UK, and ESL and English L1 students were requested to share copies of their dissertations for research purposes.
- Finally, snowball sampling was used in cases where there was low participation in each of the above methods. In this stage, the participants were asked to re-distribute the social media messages and emails to other prospective students.

Students who participated in this research shared certain demographic information to be used as variables for data classification. The variables include L1 (i.e., mother tongue or the native language) and subsequent languages (i.e., L2, L3, etc), gender, university/institute, the field of study/sub-discipline of applied linguistics, age at the time of writing the dissertation, nationality, and whether they studied their MA courses and submitted their MA dissertations in an EFL or ESL setting (the definition of each setting was given as a guide). Social and ethnic representativeness and a range of L1 backgrounds were also considered in the data collection process.

A total of 210 dissertations were collected, 70 in each of the EFL, ESL, and English L1 (labelled as ‘NS’ in tables and graphs) groups. The number of female and male participants is 48 and 22 for the EFL group, 36 and 34 for the ESL group, and 25 and 45 for the NS group respectively.

The EFL students were all Iranian nationals with various L1 backgrounds (e.g., Farsi, Iranian Turkish, Baluchi, Kurdish, Lori, Gilaki, Arabic, etc) from different geographical regions and ethnic backgrounds. They have all studied and submitted their MA dissertations in various universities in Iran with a centralised curriculum. The ESL students have different L1 backgrounds (e.g., Arabic, Japanese, Polish, Korean, Chinese, Spanish, etc) and have studied and submitted their dissertations in various universities in the UK as part of academic immersion programmes whereby the students live in the UK during and/or several years prior to the postgraduate studies. The English L1 students are all British nationals (with English as their L1/mother tongue, born and raised mainly in the UK, but some had parents with other ethnicities). They all have studied and submitted their dissertations in various UK universities. Table 5.1 shows the distribution of L1 and ethnic backgrounds for the two groups.

Table 5.1. The distribution of L1 and ethnic backgrounds of EFL and ESL participants

EFL	Farsi	Iranian Turkish	Iranian Kurdish	Lori	Baluchi	Gilaki	Iranian Arabic	Total
Count	48	12	4	2	1	1	2	70
ESL	Arabic	Japanese	Chinese	Korean	Polish	German	Other*	Total
Count	15	35	7	4	3	2	4	70

*The Other category in the ESL group consists of Singaporean, Swiss, Iranian and Mexican each with 1 participant and a total of 4 participants for this category.

5.2.2. Corpus Representativeness

Given the limited number of dissertations available, care has been taken to satisfy representativeness in the corpus by choosing similar numbers of dissertations belonging to each sub-discipline of applied linguistics for each of the three groups of EFL, ESL, and English L1 from the initial data collection files. The sub-disciplines of applied linguistics represented in the corpus are mainly TEFL/TESOL/ELT, first/second language acquisition, discourse analysis and corpus-based studies (text and corpus linguistics), and linguistics (phonology, phonetics, lexical and syntactic studies, etc); a few dissertations on sociolinguistics and cognitive linguistics with similar subject areas across the three groups were also included. Table 5.2 presents the distribution of various sub-disciplines of applied linguistics across the three groups.

Table 5.2. The distribution of sub-disciplines of applied linguistics across groups

	TEFL/ ELT	First/Second Language Acquisition	Discourse Analysis	Corpus-based Studies	Linguistics	Socio-linguistics	Cognitive Linguistics
EFL	25	18	14	7	3	2	1
ESL	22	22	7	10	4	3	2
English L1	19	15	18	13	1	3	1

However, I am not specifying these sub-disciplines as independent variables in this study. One main reason is that the above-mentioned categories of sub-disciplines of applied linguistics are based on the dominant part of the studies and/or the writers' opinion about the category. In practice, however, many studies are interdisciplinary or cross-disciplinary in nature and therefore, taking the sub-disciplinary variation in this study as a variable would have only complicated the models (e.g., refer to the discussions on the results in chapter six) and would not have contributed to our understanding of proficiency differences of English L1 vs. L2 texts further.

5.2.3. Dividing the Texts into Rhetorical Sections as Sub-genres of Dissertations

Throughout chapter three, it was established via a synthesis of many research on lexical and syntactic complexity that the values of these indices are dependent on the genre (and sub-genres), task types, and rhetorical expectations of texts (e.g., Beers & Nagy, 2011; Grobe, 1981; Olinghouse & Wilson, 2013). Therefore, in this study, I divided the dissertations into six rhetorical sections as distinct sub-genres based on their functions and the rhetorical and genre moves (e.g., move analysis), and the recommendations of previous studies including studies of linguistic characteristics of thesis/dissertation and research article sections, specifically in applied linguistics. As explained in detail throughout the literature review in chapter four, the organisational pattern of theses/research articles specified as Introduction, Literature review, Methods, Results, and Discussion (e.g., Bitchener, 2010; Bunton, 1998; Swales, 2004; Thompson, 2016) was used as the basis for dividing the texts into sections. In chapter four, I also discussed in detail the reasons for the inclusion of the abstract and conclusion sections in this model. In light of the research synthesis and conclusions of Swales (2004) who considers MA dissertations as a distinct 'genre', I consider the following rhetorical sections as sub-genres of MA dissertations and aim to examine their most prominent lexical and syntactic characteristics (e.g., in terms of the overall constructs) as will be discussed in later parts of this chapter. This division (based on the six main sections) also reflects the research synthesis studies and guidebooks mentioned in Bunton (1998, p. 40) based on the rhetorical moves and the communicative goal of the MA dissertation genre.

1. Abstract: A clear heading of 'Abstract' in the initial pages,

2. Introduction: The motivation/rationale for the study, the statement of the problem, primary research questions, a brief overview and/or background (e.g., brief description of previous related research and an evaluation of the present research), the significance of the study and/or a general contribution to the field, the definition of key terms, and the structure of the dissertation,

3. Literature Review: Extensive and detailed surveys of the subject matter of the study, establishing the main gap(s) and the statements which indicate how the gap is to be filled with the research,

4. Methodology, Method and Design: The methodological issues including the critical evaluation of/rationale for the choice of techniques and methods, the research/study design, research questions, methods of data collection and analysis, and sometimes the statistical procedures (when the statistical procedures are part of/directly accompany the results, they are considered as part of the next section),

5. Results and Discussion: The presentation of the main results along with the interpretations and relevant discussions usually accompanied by tables and graphical representations

6. Conclusion: General and specific conclusive remarks, contributions of the study, theoretical and/or research and/or pedagogical implications, limitations and delimitations, a summary of the research and findings, and suggestions for further research.

To identify the sections, the headings were used in conjunction with the content of the sections/chapters, but the contents (corresponding to the specific rhetorical functions and moves as discussed in chapter four) were given precedence in the cases where the dissertations did not follow this strict design or where the headings were not clear and/or indicative of the above categories. Regarding the cases where a student did not clearly mark the boundaries of sections, or where two or more sections were merged into one large chapter, I thoroughly studied the content and decided the boundaries based on the above criteria. A few cases remained unresolved and thus the corresponding dissertations were removed from the corpus.

Several collected dissertations had to be discarded as they did not meet the criteria for inclusion in this study (e.g., organisational patterns, lack of demographic information about the writer, English learning setting, etc); they were replaced with other dissertations that met the criteria.

5.2.4. Text Preparation and Text Cleaning Methods

A standard dissertation includes many sections and elements (commonly known as noise elements) which do not integrate with the main body of the text and whose presence could negatively affect the outcome of such analyses as this research (i.e., disfigure the real picture of linguistic profiles of texts). During the text cleaning process, these noise elements were removed and/or fixed via automatic tools or manually. The following items are instances of such noise elements in this study's corpus based on the objectives of this research:

- Quoted texts (usually specified by single or double quotation marks) which are not originally written by the students, such as embedded and block quotations,
- Instances of written/transcribed samples of the research participants that the dissertation writer illustrates,
- Non-English texts such as translated texts commonly found in discourse analysis studies,
- The 'References' and 'Appendices' sections of dissertations, as well as citations of references within parentheses in the main body of texts,
- Unstructured data and non-textual elements such as stand-alone numbers, formulas, symbols/characters, page numbers, headers,
- Tables and graphical representations such as graphs, charts, diagrams,
- URLs and illegible hyperlinks

Since this study investigates the productive linguistic knowledge of students (e.g., the production of lexical and syntactic structures) as opposed to the receptive knowledge, only the parts of the texts that were originally written/produced by the students were selected for the analysis and therefore, instances of quoted texts, tables, and other data re-produced from other scholars' works were not included in the texts for analysis.

Deleting direct quotations was necessary for obtaining an accurate frequency count of types and tokens produced by students, as well as obtaining the values of lexical measures (section 5.3). Regarding the syntactic analysis, deleting quotes which are embedded in a structure (e.g., in an original sentence written by the student) results in incomplete or fragment sentences/clauses. This issue has been considered in the development of the *Syntactic Complexity Analyzer* (*L2SCA* version 2014-01-04; Lu 2010) in which the system allows clauses "to include sentence fragments punctuated by the writer that contain no overt verb" (Lu, 2010, p. 482). The details of the ways this analyser handles the texts are given in section 5.3. The fragment part is internally/automatically tagged as FRAG and the system analyses the rest of the sentence which contains a complete phrase or clause. However, these tags including the part-of-speech tags do not affect the text analysis. It was important at this step to keep the original end-of-sentence punctuation used by the writer intact after deleting the contents of the direct quotation signalled by single or double quotation marks depending on the writing style used by the students. This ensures that the system counts the sentence fragments as a sentence. The following examples extracted from the corpus illustrate how I dealt with the direct quotations embedded in a sentence:

Example 1:
Original sentence: This finding lends support to Warschauer and Grime's statement that "the need for sensitive human readers will not disappear, no matter how closely automated scores approximate scores by expert human graders" (2006, p. 34).
Deleted part: "the need for sensitive human readers will not disappear, no matter how closely automated scores approximate scores by expert human graders" (2006, p. 34)
Kept in text: This finding lends support to Warschauer and Grime's statement that.

Example 2:
Original sentence: Rivers (1981) argued "it would be impossible to learn a language without vocabulary" (p 469).
Deleted part: "it would be impossible to learn a language without vocabulary" (p 469)
Kept in text: Rivers (1981) argued.

This approach helped me to analyse only the texts written by the students themselves which accurately shows the results of the productive lexical and syntactic knowledge of students. In the first example, the system recognises the syntactic structures up to the word 'that'; this last word will be counted as a token for the lexical analysis but the incomplete part after it will not cause any error for the syntactic analysis. In deleting the quoted texts, I followed a principle of minimal interference with the text. However, in only several instances where the deleted quoted text resulted in an unintentional creation of a new/different syntactic structure which was not written and intended by the writer, I deleted a word or two and inserted appropriate punctuation to keep the original intended structure intact. A few complicated and unresolved sentences were totally omitted. Care has been taken to minimise the instances that deleting parts of a structure could change the intended syntactic structure by the writer, but I acknowledge this possibility.

Block quotations were usually represented as extended quoted texts separated from the main paragraph by double tabs, with or without single/double quotation marks. Block quotations were relatively easy to identify.

Embedded, full-sentence, and block direct quotes were located and removed via several methods depending on the writing style of that dissertation (e.g., APA, MLA, Harvard, etc), whether they were presented with or without quotation marks, and the use of single or double quotation marks. These methods include the *re* module (regular expressions in Python's standard library) and other built-in functions in *Python (version. 3.6.3), R* statistical programme (versions 3.3.3 and 3.5.1), and manually re-checking the outcomes to remove inconsistencies and manually deleting the cases where the students did not follow any of the above rules to cite.

Single-unit words with single or double quotation marks were used for emphasis, as a direct quote, or as the field's terminology sometimes presented in the form of acronyms. These items were kept in the texts intact primarily because there were few instances of direct quotes which do not make a significant contribution to the overall values of lexical units and measures. The presence of ambiguous cases also made it difficult to understand whether they are quoted or used for emphasis. The double quotation marks around a phrase or multi-unit word which were used as emphasis and not as indicators of quotation were replaced manually with single quotation marks. The cases where one double quotation mark was used at the beginning of a quoted text and not at the end of it, were resolved manually by inserting an end-quote double quotation mark prior to the automatic deletion. There were only a few instances of quoted texts which were marked by single quotation marks. These instances were spotted in the initial eyeballing process, and the single quotation marks were replaced by double quotation marks manually so that the R and Python code can correctly detect them as instances of quoted texts. Apart from these legitimate quoted texts, I found some instances where the students used double quotation marks around their own statements as an emphasis or to mark them as the exact re-statement of previously used sentences/text such as around the research hypotheses and re-statements of results with the same wording. I omitted the double quotation marks in these cases.

The References and Appendices sections of the dissertations (Cf. 5.2.4) were not included in the sub-genres for the analysis, neither the graphics (any graphical representation such as charts, graphs, pictures which contained texts, etc). Deleting the latter instances proved to be an effortless process: converting word.doc files to .txt files automatically gets rid of any type of graphics.

Citations of references within parentheses were also deleted with the same principle that the ‘References’ sections were omitted: they not only do not add value to the lexical and syntactic profiles of the texts, but also the repetition of the same reference throughout the text negatively affects the TTR (Type-Token Ratio) values, especially when the proper names are dictionary names as well. The following examples extracted from the dissertations show various scenarios where the references were omitted or kept in the texts.

– These three types of references were deleted:

Example 1:

Original sentence: Fairclough criticised some ‘goal driven’ views of discourse as being too neglectful of a true understanding of ideology (Fairclough 1995: 45-46).

Deleted part: (Fairclough 1995: 45-46)

Kept in text: Fairclough criticised some ‘goal driven’ views of discourse as being too neglectful of a true understanding of ideology.

Example 2:

Original sentence: In the foreword of the AECC by John Healy, then Minister for Adult Skills, the AECC and Skills for Life are clearly bound together (DfES, 2001; Cooke, 2006:appendix 2 extract 1).

Deleted part: (DfES, 2001; Cooke, 2006: appendix 2 extract 1)

Kept in text: In the foreword of the AECC by John Healy, then Minister for Adult Skills, the AECC and Skills for Life are clearly bound together.

Example 3:

Original sentence: Consequently, hagwon teachers being unclear about their purpose may be why the first years of ESL/EFL teaching can be quite difficult (Brannan & Bleistein, 2012: 519 citing Warford & Reeves, 2003) and many NESTs quit early in their careers (Farrell, 2012: 436).

Deleted parts: (Brannan & Bleistein, 2012: 519 citing Warford & Reeves, 2003) (Farrell, 2012: 436)

Kept in text: Consequently, hagwon teachers being unclear about their purpose may be why the first years of ESL/EFL teaching can be quite difficult and many NESTs quit early in their careers.

--These three types of references (e.g., the integrated parts of the sentences like subjects) were kept intact in the texts:

Example 1: (Cooke , 2006) confirms that the Moser Report in 1999 was responsible for linking the Skills for Life curriculum.

Example 2: I am interested in this because bodies and individuals such as the DfES (2001), Rosenberg (2007) and Ward (2007) are keen to point out that the AECC is not intended to be used uncritically and absolutely faithfully .

Example 3: Fairclough (1992:209), suggests that a fixation on skills restricts personality, Widdowson (1990:63) points to the limitations of 'performance objectives' which correspond to the idea of the 'functional syllabus'.

The decision to include or delete tables, however, was not a straightforward one. The tables representing data analysis results as well as the tables of summaries and/or re-production of other scholars' works were deleted manually. The former items included stand-alone numerical values which were unintegrated non-textual elements and the latter items were not originally produced by the students. However, the tables which included summaries/notes of the students' own research containing texts were left intact and were included for the analysis. To ease the process, the table grids were removed to keep the contents as the main text.

Non-English texts (e.g., translations and original examples), as well as instances of samples written by the research participants in each dissertation, were identified and removed manually.

All instances of unstructured data and non-textual elements such as stand-alone numbers and formulas, mathematical and non-standard symbols and codes (e.g., coding systems introduced by the students), unnecessary parentheses, page numbers and headers as well as URLs, email addresses and illegible hyperlinks were identified and removed. The above instances which were integrated parts of the sentences and which were necessary for coherent syntactic structures were kept intact.

Apart from the above elements which were removed, I fixed the following instances:

- Missing ending punctuations: the syntactic analyser used in this study counts sentences only when ends with an ending punctuation mark.
- Double punctuation: only the appropriate punctuation based on the content was kept.
- Typographical mistakes and misspelled words: Since postgraduate students are advanced-level English writers, misspelled words were considered as typographical mistakes and not errors. Their instances were identified and corrected for accurate frequency counts and lexical analysis results.
- Numbers to words: the numerical values in the texts which formed a meaningful unit as part of the sentence were kept in the text and replaced with their alphabetic equivalents. Decimal values (61.23) were kept intact. Other numerical values such as stand-alone numbers, dates (e.g., Author, 1998 or ibid., 2011), and page numbers (e.g., p.108) were deleted.
- Symbols to words: symbols and non-textual characters which did not contribute to the meaning and integrity of the text were deleted and the rest were replaced with their alphabetic equivalents.
- Contractions: contracted forms were replaced with non-contracted forms to disambiguate the cases for POS tagging.
- Split attached words: words which by mistake are joined in the texts, such as Thelanguage, are split to their detached forms, such as The language.
- Extra white space: extra white spaces between characters, words, and between words and punctuation marks were removed.
- The bullet points and numbered lists (just the points and numbers themselves) were also omitted to avoid an unnecessary increase of tokens; this includes the numbers as the heading markers and sub-sections (e.g., 2.3).
- The dot/period after et al. (e.g., ‘Robinson et al. (1999) maintain that’ became ‘Robinson et al (1999) maintain that’), as a symbol indicating multiplication (e.g., 4 . 2), and after contracted forms or certain acronyms (e.g., Dr./Mrs./Ph.D.) were also deleted so that the syntactic analyser does not confuse it with the end sentence marker. However, one period (an end-of-the-line mark) was manually inserted after each heading/title of the sub-sections so that the syntactic analyser does not count that heading as the beginning of the next sentence.

The text checking and cleaning processes were carried out via the *re* module in *Python* (version 3.7.1), the *tm* package (version 0.7-5, Feinerer & Hornik, 2018), *stringr* package (version 1.3.1, Wickham, 2018) and *qdap* package (version 2.3.0, Rinker, 2017), and *textclean* package (version 0.9.3, Rinker, 2018) in *R* (version 3.3.3 and 3.5.1) as well as manual deletion and fixing.

It ought to be mentioned that I kept certain elements in the texts as they do not negatively affect the text analysis process. These include titles and headings, field-specific acronyms, footnotes, endnotes, and numbered lists.

5.2.5. Coding and Formatting the Files

I used alphanumeric characters to name each text file to make the process of retrieving instances easier. The file naming code (e.g., AbEFLf02) consists of the sub-genre, gender, and group initials as well as a special number for each student. The full names and demographic data for each student were sorted accordingly in separate confidential files for retrieval purposes. Finally, to make the texts legible for the analysers, I converted all texts in each section to .txt format (UTF-8 encoding).

5.2.6. Word Counts and Text Truncating

Once the texts of all dissertations for the three groups were classified into separate folders representing each sub-genre, I calculated the total word count of each folder using the *AntConc* word tokens count (version. 3.5.7 for Linux, Anthony, 2018) using the Word List tab. Unlike other parsers such as the Stanford parser and Tree Tagger which inflate the token count by considering punctuations, symbols, etc. as instances of tokens in the tokenisation process, AntConc gives a more accurate picture of the actual number of words delimited by white spaces.

Following the word counts, I truncated the texts in folders to the minimum word count of any of the three groups in each sub-genre. This process assures that the total number of words as counted by *AntConc* for every group in each sub-genre is the same/similar. This step was vital for the validity of the results and the correct comparison of lexical and syntactic analyses across the groups. The approximate total token count of the three groups is 15,400 for the Abstract, 81, 600 for the Introduction, 339,400 for the Literature Review, 177,400 for the Method and Design, 302,500 for the Results and Discussion, and 106,300 for the Conclusion sub-genres of the dissertations. The total token count of the entire corpus is 3,069,192.

5.2.7. An Exploratory Pilot Study

Following the research design specifications as explained in 5.1 and 5.2, a pilot study was carried out with a sub-set of dissertation abstracts that have been collected up to that point. The piloting phase was conducted to assess the feasibility of the research methods, and to examine the effectiveness of the selected indices up to that point to guide the final measure-selection phase of the final study as will be discussed in the next section.

Since this phase was exploratory in nature, I used all the indices available in Lexical Complexity Analyzer (LCA) and Syntactic Complexity Analyzer (L2SCA); a more detailed description of these analysers will be presented in 5.3.2. Several previous studies have also used all the indices in these two analysers to explore and examine the effective measures of proficiency in various SLA corpora (e.g., Kim, 2014; Lu, 2010, 2012; Lu & Ai, 2015).

The data consisted of a total of 150 abstracts, 50 abstracts from each group. This was the amount of data that was collected and processed (e.g., text cleaning and pre-processing) up to that point. The decision of including the abstracts was also motivated by a comparison of the results of the effective measures with previous studies, as most of the previous research in this area has been conducted on relatively short texts (e.g., argumentative essays). A comparison of these results with the final study, i.e., pilot study with 150 abstracts vs. the main study with 210 abstracts, also informs the effect of sample size on the number and type of measures that show significant between-group differences.

This phase was followed up by the analysis of the entire corpus, i.e., genre-aggregated lexical and syntactic datasets, for an informed decision on the selection of the final set of measures (discussed in the following sections) prior to analysing separate rhetorical sections. The statistical procedures and results of the pilot study and both aggregated datasets will be provided in detail in chapter six, section 6.2.

5.3. Data Analysis Procedure: The Measure-selection Process, Quantification Methods, and Analysing Programmes

In the following sub-sections, I give detailed accounts of the two sets of measures, along with the description of the analysis methods and tools/programmes to obtain quantitative values. Based on the synthesis of previous research cited throughout the literature review chapters as well as the section 5.3.1, I selected two sets of lexical and syntactic complexity indices which are relevant, sufficient, and necessary based on the objectives of this study and based on the research design. As explained in sections 5.1.1 and 5.1.3, apart from finding the differences between English L1 and L2 (EFL and ESL) groups, in the present study I also intend to test the performances of similarly-calculated vs. dissimilar lexical and syntactic proficiency measures (how well they distinguish between the three groups with different English language backgrounds and academic setting) and to find how these measures relate to each other in the six rhetorical sections. I will also use statistical modelling methods to see how well we can predict group membership based on the complexity measures and which variables are most predictive of these groups.

5.3.1. The Measure-selection Process

In chapter one, section 1.1.4, I elaborated on the usefulness of NLP tools that can automatically analyse a set of features in large corpora and cited scholars like Doró and Pietilä (2015) who listed the studies that show high correlations between the scores from these analysers and human judgments of writing quality, etc. I also discussed that despite these advantages, additional issues have been acknowledged. One is the availability of a multitude of measures, some of which have not been thoroughly tested across corpora of different genres, proficiency levels, etc. Many of these indices, e.g., lexical diversity indices, are different adaptations of the same quantification methods that are often proposed by quantitative linguists and writing assessment researchers as alternatives to the simple TTR and to reduce the effect of text-length dependency of various measures. Other indices simply gauge complexification of texts from different angles that could be more suitable for a specific type of research. As one objective of this study, I aim to test two large sets of lexical and syntactic complexity measures to arrive at two small sets of unique measures that are the most effective ones at capturing text complexity differences of postgraduate academic writing in various rhetorical sections of MA dissertations written by students with different English language backgrounds. Various statistical tests in chapter six will be employed to answer the research questions specified in chapter one in the measure-testing process. To be able to carry out this measure-testing process, I first selected pairs or groups of similarly-calculated indices (e.g., different quantification methods for capturing the same underlying construct) that are proposed and used in the literature. Sections 5.3.1.1 and 5.3.1.2 will provide detailed descriptions of these measures in each pair/group. In each pair or group of measures at least one measure has been already validated and/or reliability-tested based on the many studies that I have reviewed in chapter three. I have already elucidated the criteria of reliability and criterion validity (e.g., concurrent and predictive validity) in section 3.3 and have synthesised the collective evidence for them throughout chapter three. This will be taken as the main criterion for selecting the main measures in the pairs/groups, especially various lexical diversity measures. These studies provide evidence that any single or set of these measures are indicative of linguistic proficiency (differences), descriptive of distinguished language performance, e.g., between English L1 and L2, and suggestive of language development and progress at the postgraduate level in academic writing, or at the advanced level EFL/ESL writing and speaking. The performance of the rest of the measures in each pair/group will be compared to the already-validated and reliable measure(s), i.e., a comparison method based on concurrent validity (a sub-category of criterion validity). This will be carried out via various statistical methods in the next chapter, e.g., their effect sizes in capturing text differences of the groups. I will further examine whether the various measures that have been used as representative indices of a certain construct in previous studies, do actually load together, e.g., based on the exploratory factor analyses in chapter six.

Among these measures in each group, a few have also shown inconsistent findings by various scholars (e.g., subordination indices in Yoon, 2017 vs. Ai & Lu, 2013 vs. Paquot, 2019); these measures have also been reviewed in chapter three. I hypothesise that these measures could be more suitable for a specific type of data, e.g., general vs specialised writing corpora, or proficiency levels. Comparing the results of these measures in this study's specialised corpus (e.g., in various rhetorical sections by different groups) with more general-purpose corpora in SLA studies could be informative in this regard.

Additionally, the measures with the same denominator in their formulas will be compared. For example, the comparison of CT/T, CN/T, DC/T, VP/T, and CP/T informs the variation in dependent clauses vs. complex nominals vs. coordinate phrases in T-units across groups and rhetorical sections. Similar is the comparison of syntactic indices with clause as the denominator, or the lexical measures of lv, vv2, nv, and adjv with N_{lex} (lexical tokens) as the denominator.

The two measures of lexical density and RTTR will remain as separate indices which do not belong to any group. The performance of these isolated indices in this study will be compared to previous studies with different research designs to examine their effectiveness in capturing English L1 vs. L2 text differences at different proficiency levels.

Finally, the results of the pilot study were considered as an additional reason for selecting the final set of measures. Since the pilot study was carried out with relatively short texts of abstracts, care has been taken to use the evidence as an additional reason (e.g., beside the validity and reliability reports and the recommendations of previous studies), in cases where a specific measure has been shown inconsistent results in the related literature, or in cases where a measure has not been used in advanced English writing samples.

5.3.1.1. Lexical Complexity Measures and their Quantification Methods

A set of 22 lexical density, lexical diversity/variation, and lexical sophistication indices were selected based on the above criteria and with specifications in the following paragraphs. Each index measures its representative construct from a different angle and/or with a different quantification method. When T and N are stated in the same formula, T refers to the number of types and N refers to the number of tokens (see the related literature reviewed in chapter two for a table of all formulas).

Both Lu (2012) and Šišková (2012) found weak correlations between these three constructs of lexical complexity which means they indeed are distinct constructs. Construct-distinctiveness of these categories will be further examined based on this study's corpus of postgraduate academic writing in chapter six. The relatively high correlations among the measures in the same construct or sub-construct and lower correlations of these measures with indices from another construct in Lu (2012) is evidence of construct validity of the measures that are analysed in LCA and used in this study as well as in LCA-AW (apart from sophisticated measures in LCA-AW).

With respect to the three constructs that will be explained in detail in the following sections, one might argue that the presence of field-specific terminology can affect different lexical complexity values which might not necessarily be related to linguistic proficiency. However, there are three points to be considered:

1. The field-specific terminology is present in every sub-discipline of applied linguistics; therefore, it is expected that all students at the postgraduate level to use such instances in dissertations.
2. This study analyses groups performances rather than individual production; in this regard, the low and high numbers of such instances in individual dissertations could balance each other out when taken as a group.
3. Such field-specific terms will be inevitably repeated throughout each text; this consequently can lower lexical diversity values just as any other word (high-frequency words) in the text.

Lexical Density:

N_{lex} / N (Ure, 1971) or the ratio of the number of lexical items to the total number of words/tokens is regarded by Halliday (1985, p. 62) as “the kind of complexity that is typical of written language”. Likewise, a “highly information-packed, lexically dense” (Halliday, 1989, p.87) type of writing is regarded as “a major source of syntactic complexity of academic texts” (Liu & Li, 2016, p. 51). Read (2000) similarly proposes that lexical density is a characteristic of written language where texts represent a more concentrated proportion of lexical items in the form of information and ideas. Among text types, the density of formal texts like academic writing is shown to be higher than that of informal texts (e.g., in Biber 2006). As an instance, the process of nominalisation reduces the grammatical words and contributes to higher lexical density. Consequently, the texts with these characteristics such as a dense use of nouns are more informative and can be regarded as a characteristic of the academic genre and advanced writing (Biber, 2006; Biber & Gray 2010, 2016).

This index, which is believed to show the information content of a text, showed significant between-proficiency level differences and a strong predictor of the academic writing in Kim’s (2014) study. The mean number of words in the texts in Kim’s study is less than 310 words for the advanced group and, therefore, these texts are classed as relatively short texts. In Gregori-Signes and Clavel-Arroitia’s (2015) study, students at higher levels of proficiency produced texts with higher lexical density values. Linnarud (1975, cited in Arnaud, 1984) also found that the lexical density of English texts written by English native speakers has higher values than those written by EFL speakers. As a token-token ratio, this index is also assumed not to be affected by sample size (Malvern et al., 2004). Therefore, I investigate whether short vs. longer texts of dissertation sub-sections could have an impact on the obtained values and whether English L1 vs. L2 texts demonstrably differ in their lexical density. An extended explanation about lexical density is provided in 2.2.1.

Lexical Sophistication:

Two measures of lexical sophistication (LS1 and LS2) and two measures of verb sophistication (VS2 and CVS1) are selected to represent and quantify this construct. Detailed discussions on the effectiveness of various lexical sophistication indices are presented in 2.2.3.

LS1 or Lexical sophistication type I was initially proposed and analysed by Linnarud (1986) and Hyltenstam (1988, cited in Lu, 2012) and used in Kim (2014) and Paquot (2019). This measure which is calculated as the ratio of sophisticated lexical tokens to the total number of lexical tokens or N_{slex} / N_{lex} , was shown in Kim (2014) to linearly increase with higher proficiency levels. LS1 only showed a small difference (linear increase) between C1 and C2 CEFR-based proficiency groups in Paquot (2019). This index was also among the measures in the pilot study which showed significant differences with large effect sizes between the English L1 students and EFL student group as well as between both non-native English student groups in their academic writing. In all these three studies sophisticated lexical words were only based on their absence from the general BNC frequency word list.

LS2 or Lexical Sophistication type II is calculated as the ratio of sophisticated types to the total number of types (T_S / T) and thus a more strict criterion than LS1 (i.e., sophisticated types in this index vs. all sophisticated lexical items in LS1). Both lexical sophistication indices showed medium to large effect sizes for the between-group differences in the pilot study. Laufer (1994), Linnarud (1986), Lu (2012), and Paquot (2019) are among the SLA studies that used this measure for analysing written and spoken corpora of English learners. While the first two studies reported significant differences between the values of this measures in written learner corpora of English L1 vs. L2, Lu’s (2012) study did not find any such differences in the oral narratives of English L1 vs L2. This could be evidence of the effect of this measure in written vs. spoken lexical proficiency as stated in Lu (ibid.) as well. In Paquot (2019), LS2 showed non-significant differences between non-adjacent proficiency levels of B2 vs. C2, as well as between C1 and C2 in a specialised SLA corpus. At the time of the final analysis of this study, no further evidence of the usefulness or otherwise of this index in measuring proficiency differences in academic writing has been reported; therefore, I included this index for the measure-testing process to see its effectiveness compared to LS1.

VS2- Verb Sophistication II- This verb-based measure which was first proposed and investigated by Chaudron and Parker (1990) was among the indices which showed a statistically significant difference between two pairs of comparison in the pilot study. It is calculated as T^2_{sverb} / N_{verb} where T is the number of types, sverb is the number of sophisticated verbs, and N_{verb} is the total number of verbs. Unlike VS2, in the VS1 (the first type of verb sophistication in the LCA analyser measured as T_{sverb} /N_{verb}) formula, the sophisticated verb types are not squared, and this might be an explanation for smaller effect sizes obtained in the pilot study. Squared T will increase the weight of sophisticated verb types in the VS2 equation against the total number of verbs (i.e., if there are too many general/ high-frequency verbs used in the text). Wolfe-Quintero et al. (1998) and Lu (2012) both commented that squaring the number of types in the formula of VS2 reduces the sample size effect. VS2 also showed non-significant differences between non-adjacent proficiency levels of B2 vs. C2, and C1 and C2 in Paquot (2019).

CVS1- Corrected VS1- This index that is calculated as $T_{sverb} / \sqrt{2N_{verb}}$ is recommended by Wolfe-Quintero et al. (1998). Both simple VS1 and CVS1 (as labelled and calculated in LCA) are other verb-based lexical sophistication measures which showed between-group differences in the pilot study, but the CVS1 produced larger effect sizes. The simple VS1 is a ratio of sophisticated verb types to the total number of verbs (as discussed in the previous measure) and showed smaller effect sizes than the corrected one. The square root of $2N_{verb}$ in the CVS1 formula reduces the weight of the total number of verbs; this helps to obtain a better value of sophisticated verbs when there are too many general/high-frequency verbs in the text. One might notice that VS2 and CVS1 are simply two similar ways of capturing the same concept, one with increasing the weight of sophisticated verbs (VS2) and one with reducing the weight of high-frequency verbs (CVS1). In this work, I intend to examine which method leads to larger effect sizes for the same group of texts. Both Wolfe-Quintero et al. (1998) and Lu (2012) believe that the VS1 transformations (e.g., VS2 and CVS1) better distinguish between proficiency levels and therefore the simple VS1 is not included in this study's set of measures. Just like the previous sophisticated indices, the values of this index also showed differences (though statistically non-significant) between non-adjacent proficiency levels of B2 vs. C2, as well as between C1 and C2 in Paquot (2019).

It is important to consider that the mentioned studies, including the pilot study of this research, used a general-purpose word list as an external basis for the frequency of lexically sophisticated items. The final analysis of the present study is the first work that considers a strict criterion for the lexical items to be regarded as sophisticated. That is, lexical sophistication indices in this study will be filtered through a general corpus word list (BNC) as well as a discipline-specific word list (BAWE list for linguistics). These word lists act as the external reference of word frequency bands; sophisticated verbs do not appear in the 2000 most frequently-used words in the BNC list, nor in the 100 most frequently-used academic writing words in the linguistics BAWE word list (see section 5.3.2). This provides a good basis to compare the effectiveness of these sophisticated indices in general SLA vs discipline-specific texts. I will also compare the performance of these measures in the pilot study abstracts using the LCA analyser (filtered via BNC word list only) vs. the final study abstracts (with both word lists). Vermeer (2000) suggests that indices which gauge the quality of words by their levels of frequency based on external reference points (frequency classes in corpora) are better predictors of vocabulary size than simple TTR-based measures (as will be discussed below). Further details about the implementation of these criteria will be provided in section 5.3.2.

Lexical Diversity/Variation:

A simple variation of type-token ratio (TTR) referred to as RTTR or root ttr, as well as five measures based on the word strings and segments (ndwerz, ndwesz, msttr, mattr, mtld) together with the two variations of the D measure (vocd-D, HD-D) based on word samples, three logarithm-based measures (logttr, Uber, maas), and six lexical variation indices based on the TTR of word classes (lv, vv1, cvv1, vv2, nv, adjv) are selected to represent lexical diversity or variation in this study. Using various statistical tests, I will examine the effectiveness of each measure in these groups/pair of related and similarly-calculated measures in capturing lexical diversity differences in various rhetorical sections of the dissertations by the three groups. In-depth discussions of these tests and the usefulness of each measure will then be presented throughout chapter six.

TTR or ***Root TTR***, also known as the ***Index of Guiraud*** (Guiraud, 1954) is a mathematical transformation of TTR and is computed as T / √N where T is the number of types. Treffers-Daller (2013) indicates that this index compensates for the issue of text length to some extent. As Daller, vanHout and Treffers-Daller (2003) mentioned, the square root in the denominator results in a higher value for longer texts with the same TTR as shorter texts, and thus maintains the same TTR for a longer text. However, Jarvis (2002) concludes that this measure is not a good model for the actual TTR curves and therefore, still remains text-length dependent. This discrepancy could be due to the fact that the texts in Jarvis's study were shorter than 500 words. In Lu's (2012) study of oral narratives, RTTR was shown to have a high correlation with test takers' rankings; it also showed significant between-proficiency level differences. High correlations between RTTR and other indices of lexical diversity in Lu (2012) is also evidence of its construct validity. Kim (2014) also showed that the values of RTTR increase with the increase in writing proficiency; it further discriminated between different proficiency groups' writings. Paquot's (2019) analysis of a specialised academic writing corpus also shows the values of this measure increase linearly across B2, C1, and C2 proficiency levels.

Lexical Diversity Indices Based on Word Segments or Samples:

The measures in this group of lexical diversity have been more studied compared to other groups/pairs of diversity measures. Among the measures in this group, the two measures of MTLD and Vocd-D are extensively studied and reliability- and validity-tested as will be discussed. The values of the rest of indices in this group will be compared with these two strong measures as part of the process of construct validation that has been discussed in previous sections.

NDW is the ***Number of Different Words*** used in a language sample. This measure is usually counted from a baseline of utterances (e.g., 50 utterances in Klee, 1992 and 100 complete utterances in Miller, 1991). Even though this measure did not show any between-group differences in the academic writing of the three groups in the pilot study, significant differences were obtained in the studies of Kim (2014) as a predictor of L2 academic writing proficiency and Lu (2012) as the quality of transcribed oral narratives of college students. However, the simple NDW index is suggested to be text-length dependent and several scholars (e.g., Malvern et al., 2004; Lu, 2012) recommended standardised indices which select subsamples of the same size instead. Among the four indices based on the NDW in Lu's 2012 classification, the ***NDWERZ*** and ***NDWESZ*** measures are used in this study; they are recommended by Malvern et al. (2004) and Lu (2012) as standardised indices that, unlike the simple number of types, are not affected by text length. Both indices select 10 random sub-samples of 50 words to get the averages of NDW (the numbers are set to comply with the LCA-AW analyser, cf. 5.3.2). The sub-samples of the former measure include a random but standard number of words from the sample and the latter measure's sub-samples include a standard number of consecutive words, but the starting point is randomly selected. There were high correlations between these two measures and the test takers' rankings as well as significant differences between proficiency levels in Lu's (2012) study. No study so far has reported the effectiveness of any of the two measures over the other in capturing differences of advanced-level English texts by English L1 vs. L2 students. Therefore, I will investigate the usefulness of either of the quantification methods as well as comparing them with the rest of the diversity measures based on word segments and samples.

MSTTR- or the ***Mean Segmental TTR***, (Johnson, 1944) is a type-token based measure of lexical variation which averages the TTRs from all fixed-size segments of the texts and thus overcoming the problem of differences of sample sizes (see the discussions in Jarvis, 2013 for instance). In this study, the texts are divided into segments of 50 words. Torruella and Capsada, (2013) found out that MSTTR is one of the indices that functions independent of text-length based on the method of equal and cumulative blocks of words. Johnson (1944) suggested this measure as a solution to the sample-size dependency problem and McCarthy and Jarvis (2010) show how this measure works better with longer texts. In Lu's (2012) study of oral narratives, this measure is found to have a high correlation with test takers' rankings and there were significant between-proficiency level differences based on this measure as well. Jarvis (2013) has tested a variety of lexical diversity measures and found that the msttr index holds the "same size constant while calculating the mean TTR across different segments of a text" (p. 94). In this study, the effectiveness of this measure will be compared both with the rest of lexical diversity measures based on the word strings/segments, as well as the values of the same measures in different rhetorical sections.

MATTR or the ***Moving-Average Type-Token Ratio*** (Covington & McFall, 2010) computes the lexical diversity of a text by assigning a moving window and estimates the type-token ratios for fixed-length successive windows. For example, if a 10-word window is selected, it estimates the TTR for words 1-10, then for the words 2-11, and then 3-12 and so on till the whole text is covered at which point the final score is the average TTR estimates. This moving window is a feature which distinguishes it from the MSTTR measure and which allows the words in a text to be successively calculated, not just fixed successive chunks or segments. Furthermore, unlike MSTTR, using MATTR does not result in data loss. Jarvis (2018, personal communication) recommended this measure as he believes it is not affected by variations in text length. Jarvis (2002) also believes that sequential selection of samples (e.g., in msttr and mattr) is a better method for measuring texts than a complete random sampling of words. The effectiveness of either of the two approaches in capturing differences in academic writing of English L1 vs. L2 texts will be further examined throughout chapter six. Their values will also be compared with the rest of the indices in this groups across groups of students and rhetorical sections.

MTLD- or the ***Measure of Textual Lexical Diversity*** was first proposed and reliability-tested by McCarthy (2005), and tested and validated by McCarthy and Jarvis (2010). Koizumi (2012) also reported the validity of this measure. MTLD is computed as "the mean length of word strings that maintain a criterion level of lexical variation" (McCarthy and Jarvis, 2010, p. 381), e.g., maintaining a type-token ratio of 0.72 in the mentioned study. This means that it measures the TTR in a sentence till the ratio falls to 0.72, at which point one factor is counted and the TTR is re-calculated for the rest of the string. The MTLD value is then calculated as the ratio of the total number of words to the total number of factors. This process is carried out for a forward and backward analysis of the text for an average of the outcome, which is the final value of MTLD. It is reported that MTLD is robust to variations in text length for a word range of 100-2000 (Crossley et al., 2009; McCarthy, 2005; Torruella & Capsada, 2013) as well as to variations in sample size in general (Jarvis, 2013; McCarthy & Jarvis, 2010) and thus a suitable candidate to analyse texts of this study’s corpus, e.g., for the comparison of the values of this measure across rhetorical sections with various sample sizes. The effectiveness of this index compared to MSTTR will be particularly examined. This is because Jarvis (2013) indicates that these two measures.

> “are essentially mirror images of each other. MSTTR holds the sample size constant while calculating the mean TTR across different segments of a text, whereas MTLD holds TTR constant (usually at .72) while calculating the average number of words in any segment of text that remains above the TTR cutoff value” (p. 94).

Vocd-D- The D measure was first proposed by Malvern and Richards (1997, 2000) to obtain the lexical variation of the whole text by capturing the rate of decrease of TTR by finding the best-fitting curve of TTR. It is calculated as TTR = (2/DN) [(1 + DN) 1/2 – 1]; for more details of how D is calculated using this formula see McCarthy and Jarvis (2007). Higher token curves are the results of greater diversity in the text; in other words, the value of D specifies the height of the curve and hence the value of lexical diversity. Curve fitting is a process of using a mathematical function which can fit all or a specified number of data points (e.g., on a curve) in its best possible way. This measure then has undergone more developments by Malvern, Richards, Chipere and Duran (2004) and eventually, the *vocd* program was developed by McKee (McKee et al., 2000) based on the new method that uses the random sampling (without replacement) which is somewhat different from the original method. The updated version is called the adapted D or vocd-D (hereafter labelled as ‘vocd’ in this thesis). This probabilistic method randomises the order of words before measuring the TTR segment so that every word has the chance of being included in the calculation. This randomisation is repeated 100 times to obtain a mean TTR for 35-token samples and 100 times for 36 to 50-token samples. Final D values (after repeating this procedure three times) range between 10 and 100 where the higher values indicate greater lexical variation. McCarthy and Jarvis (2007) examined the text-length dependency of Vocd and found that texts between 100 and 400 tokens might be suitably compared; this is important for short abstract texts in the present study. The D measure is also validity and reliability-tested by Malvern et al. (2004); they also compared the mean values of this measure across corpora, e.g., adult learners, academic texts, etc. Furthermore, McKee et al. (2000) reliability-tested the Vocd-D measure using the split-half method and found that sample size did not have a significant effect on D scores. Relatively high correlations of Vocd-D with the rest of lexical diversity measures and much lower correlations with the indices of other constructs in Lu (2012) also can be taken as its construct validity.

The above studies reported that the Vocd measure strongly correlated with other measures of language which are validated as well as with some developmental variables. The Vocd measure in Lu's (2012) study of oral narratives showed significant correlation with test takers' rankings as well as significant differences between proficiency levels. Yoon (2017) also recorded this measure as a significant predictor of L2 writing proficiency. Vocd is also shown to differentiate between NS and ESL academic writing in Gonzalez (2013) but not in Pietilä (2015). Finally, McCarthy and Jarvis (2010) found that the values and the correlations of these two measures are affected by text register variations, with Vocd-D having more distinguishing power in register variations.

HD-D- This index which was proposed by McCarthy and Jarvis (2007) is a variant of the D measure (vocd being an 'estimate' of HD-D) and is obtained from the hypergeometric distribution function. This is a probability distribution which calculates the probability of the number of successes of random draws without replacement in a number of trials. HD-D calculates, "for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text" (McCarthy & Jarvis, 2010, p. 383). The sum of all the probabilities for all lexical types in the text is used as an index of the text's lexical diversity. McCarthy and Jarvis (2010, p. 384) comment that vocd and HD-D have different scales, i.e., "the HD-D output is literally sums of probabilities, whereas vocd-D output is essentially sums of probabilities converted to type-token ratios and, then again, from type-token ratios to a D value". HD-D is reported to be less affected by text length based on the method of the equal and cumulative blocks of texts (Torruella & Capsada, 2013). DeBoer (2014) analysed 1200 college-level essays of 100-400 token range from English L1 and L2 writers and found that the two indices of Vocd and HD-D have very high correlations (little/no effect of language background on the correlation values). The inclusion of this measure besides the Vocd allows me to examine this assertion and compare the obtained values for longer dissertation texts (e.g., literature review and result sections) to find out if any of them can distinguish better between the three groups. I will also examine their convergent validity (i.e., if the two theoretically-related indices of the same sub-construct are in fact related, e.g., are found out to have a high correlation).

Logarithm-based Measures of Lexical Diversity:

The three lexical diversity measures of logttr, uber, and maas are based on logarithm functions. Logarithms are the inverse of exponentiation: the exponentiation enlarges small numbers and the logarithm downsizes large numbers. Winter (2019) explains that logarithms are non-linear transformations which can transform the numbers to orders of magnitude. Using logs can be helpful if we have a non-normal distribution like the TTR of long texts. In such scenarios, we have a positive skew because as the number of tokens increases, the number of types decreases. Such non-linear transformations get rid of the skew and make the distribution more symmetric. The logarithmic transformations in this study use base 10. Among the three measures, maas has been reliability tested for the effect of text length as will be discussed. I will, therefore, compare the values of the other two indices against maas. However, the three measures use fundamentally different quantification methods based on logarithm, with uber and mass being the inverse of each other. Since these three indices have not been investigated in advanced levels of writing proficiency, I will examine which one can better distinguish between such texts from postgraduate students with different English language backgrounds. I will further examine if the three indices line up on the same factor in exploratory factor analysis in chapter six, e.g. if indices that use logarithms with fundamentally different approaches are in fact related.

LOGTTR- This Bilogarithmic type-token ratio also known as ***Herdan's C*** is measured as Log T / Log N (Herdan, 1960) where Herdan argues for the constancy of logarithmic ratio of type and tokens. According to Malvern et al. (2004, p.27), it is conceptually defined as "the rate of increase of types with increasing token count will be proportional to the TTR for any given value of N". The logarithmic functions in both LOGTTR and Uber's index (the next index discussed below) simply change the shape of the TTR curve and as Malvern et al. (2004) state, "linearise" the type-token curve. Logttr was used in Lu's (2012) study but did not reveal significant between-proficiency differences in general SLA study of narratives. However, it did show between-group differences in academic writing of English L1 vs. L2 students in the pilot study with medium to large effect sizes.

Uber's U index (Dugast, 1978) is another Log-based transformation of TTR but unlike LogTTR, it scales down the ratio of tokens to the division of tokens to types. This calculation is formulated as (log N) ^{2} / (log N – Log T). This is a notational inverse of Maas constant (cf. The Maas index below). Jarvis's (2002) found that this index provides accurate fits for TTR vs. token curves. I included this measure primarily to compare the values and the performances of these three log-based measures (LogTTR, Uber, and Maas) as well as finding the effect of longer texts (e.g., longer sub-sections of dissertations) on the obtained values. Higher U values denote more lexical diversity. This measure showed significant correlations with test takers' rankings as well as significant between-proficiency level differences in Lu (2012) study of oral narratives. Jarvis (2002) study of adolescent narrative texts (which were shorter than 500 words) also found that U accurately models the actual TTR curve and is an optimal model of lexical diversity of whole texts, as well as texts reduced to content words (their function words were omitted to check the effect of TTR curve-fitting). However, in the pilot study, LogTTR showed larger effect sizes for the between-group differences of English L1 vs. L2 than the Uber index considering that the texts (Abstract sections of dissertations in the pilot study) were shorter than 400 words. It is unclear if comparable results would appear with longer texts.

Maas- The Maas index of $\boldsymbol{a^2}$ (also its variants ***lgV0***, and $\boldsymbol{lg_eV0}$ which are not used in this study) is proposed by Maas (1972, cited in Treffers-Daller, 2013) are logarithmic corrections and the a^{2} index is calculated as (Log N – Log T) / (Log N)^{2}, where V is the number of types and N the number of tokens. Maas is the inverse of the U's notation (cf. Uber's measure above). Maas is another index reported to be text-length independent and thus recommended as a reliable index of lexical diversity (McCarthy & Jarvis, 2007, 2010; Torruella & Capsada, 2013). McCarthy and Jarvis (2007) for instance, found the Maas index to be independent of text length with four different ranges of text length. This log-corrected TTR is reversed (e.g., 1-Maas) to assign higher scores to more lexically diverse texts. While higher values for other measures denote greater lexical diversity, higher values of Maas indices denote lower lexical diversity. In Torruella and Capsada (2013) maas is also shown to have lower sensitivity to text length compared to other indices of lexical diversity.

Lexical Diversity based on Type-Token Ratio of Word Classes:

Three verb variation measures, as well as lexical variation, noun variation, and adjective variation indices, are selected based on the categorisation in Lu (2012). As mentioned before, I will investigate the four measures of vv2, lv, nv, and adjv with the same denominator (lexical tokens) to examine the extent to which the total variation in lexical types can be attributed to the noun, verb, and adjective types, and to compare the amount of verb, noun, and adjective types for any distinct pattern of any of these non-repetitious production units across groups and rhetorical sections. Among the measures in this category, lv has been reliability tested (Engber, 1995). The collective evidence from the studies on the two verb-based measures of VV1 and CVV1 also reflect repeated sampling reliability (e.g., the discussion in section 3.3) and a strong relationship between these measures and proficiency.

The three ***Verb Variation*** measures of ***VV1*** (measured as T_{verb} / $N_{verb)}$, ***CVV1*** (Corrected VV1, measured as T_{verb} / $\sqrt{2}N_{verb}$) and ***VV2*** (Verb Variation 2, measured as T_{verb} / N_{lex}) were found to significantly distinguish between English L1 vs. L2 students' academic writing in the pilot study with medium to large effect sizes. The ratio of the number of verb types to the total number of verbs showed significant differences between native and L2 French writers (Harley & King, 1989). Both VV1 and CVV1 showed between-proficiency level differences in Lu (2012) study of oral narratives by university students as well as in Mazgutova and Kormos (2015) academic writings of ESL groups. Paquot (2019) also reported that the values of CVV1 index increase across CEFR-based proficiency levels and VV2 in non-adjacent levels in the academic writings of EFL groups. These three verb-based diversity measures have verb types in their numerators but different denominators. I will examine whether any of the methods can more effectively distinguish between postgraduate academic writings of the students with different English language backgrounds.

LV or ***Lexical Word Variation*** is an index based on the type-token ratio of word classes and is measured as T_{lex} / N_{lex} or the ratio of the number of lexical word types to the number of all lexical words in a sample. Linnarud (1975) proposed this measure in the form of percentage ratio. Engber (1995) found a significant positive correlation between lexical variation values and the quality of writing of ESL students from various L1 backgrounds. In the pilot study, I found medium to large effect sizes for the differences between English L1 vs. L2 academic writing for the values of this measure. Paquot (2019) also reported that the values of LV index increased in non-adjacent proficiency levels of three groups of EFL academic writers.

NV or ***Noun Variation*** is the next measure based on the TTR of word classes and is included in this study's analysis. It is measured as T_{noun} / N_{lex} which calculates the number of noun types to the number of lexical tokens. This measure showed an incresae between the specialised academic texts of B2/C1 and C2 proficiency levels in Paquot (2019). The noun variation measure showed close to medium effect sizes for the differences between English L1 vs. L2 academic writing in the pilot study as well. Biber (2006) and Biber and Gray (2016) illustrates that written university registers heavily rely on nouns and nominalisations.

ADJV or ***Adjective Variation*** index is selected as the last lexical diversity measure based on the TTR of word classes. It is calculated as T_{adj} / N_{lex} or the number of adjective types to the number of lexical words. Paquot's (2019) results showed a non-significant increase in adjv values between specialised academic texts of B2/C1 and C2 proficiency levels. The values for this measure also showed a medium effect size for the English L1 vs. EFL comparison set in the pilot study.

It ought to be acknowledged that the above-mentioned measures have attracted some criticisms, some of which are reflected in Chapter two. However, the validity and reliability of the above-selected indices have been reported and confirmed by various scholars. The indices that are included in this study's set of measures have been previously compared with the holistic ratings of experts with high correlations between the analysers' values and the experts' ratings. Engber (1995), for instance, found a significant relationship between the value of Lexical Variation (cf. LV measure in the set of measures in this study) as well as lexical density and the holistic ratings of intermediate and advanced ESL students' writing with an inter-rater reliability of 0.93. Jarvis (2002) also found a significant correlation between the lexical diversity measures of D and U and the quality of writing (based on holistic ratings of experts) of both native and non-native English students. Jarvis (ibid.) and Arnaud (1984) also found significant correlations between various lexical diversity measures and the vocabulary test scores. In Arnaud (ibid.) lexical variation was measured as V_{lex} / N and the index of rare (sophisticated) words was measured by the rareness score (see the details in chapter two) as V_{rare} / V_{lex} where rare words do not appear on the official list of Classe de Troisieme; he randomly selected 180 words in all university essays to standardise the length of texts. He further reported high reliability of lexical richness variables after text shortening. Lu (2012) also checked the measures included in the Lexical Complexity

Analyzer (LCA, cf. 5.3.2) against test takers' rankings by expert raters as well as a re-evaluation by a quality control committee.

Among the prominent measures included in the analysers in this study (section 5.3.2), there are several indices which did not fit the purpose of this study (e.g., the specifications of the research design) and hence were left out from the set of measures in this study. Yule's K (Yule, 1944) is an instance of such indices left out since its reliability for language acquisition research (especially with shorter texts) is questioned by Jarvis (2002). Jarvis (ibid., p. 60) further maintains that this measure remains independent of text length as far as "its probability assumptions are met, i.e., when the order of the words in the text is randomized".

Even though I have truncated the texts in this study to control for the text length in individual rhetorical sections, I did not include the simple TTR measure because 1) I have already included more robust alternatives to TTR and 2) because of the extreme text-length dependency of this measure, I cannot compare the TTRs of short vs. longer texts/rhetorical sections. Consequently, I cannot compare it to other lexical diversity measures in this study that have been reported to be relatively robust to sample size (e.g., across rhetorical sections). Vermeer (2000, p. 69) calls the simple TTR "the worst measure of lexical richness". The non-standardised forms of the Number of Different Words indices, namely the simple NDW and NDWZ, were also omitted from the final set of measures (cf. the section of the Number of Different Words above). The next measure that I dropped is CTTR or Corrected TTR (Carroll, 1964). As Vermeer (2000) explains, this measure is a replicate of RTTR (Index of Guiraud), and it does not make sense to multiply the square root of tokens by the factor 2 (see the quantification method of CTTR in chapter two). Lu (2012) also found that CTTR and RTTR are perfectly correlated and thus are essentially the same measure. The results of the pilot study based on dissertation abstracts also indicated that both CTTR and RTTR have very similar effect sizes. The rest of the measures that were reviewed in chapter two had a similar measure to them that has already been selected among the set of indices in this study (e.g., Orlov's Z index and Somer's S index cited in Malvern et al., 2004, both as logarithm-based measures, or Mendelsohn's lexical variation measure, see chapter two). A few other measures (e.g., Brunet's W index, Sichel's S index, Rubet's K index, all tabulated in chapter two) were also not selected because at the time of conducting this study's analysis, I did not have any access to any programme to compute them nor did I find sufficient evidence of their usefulness as indicators of linguistic proficiency in advanced English texts.

The next index that I did not include in the final set of measures is VS1 or the verb sophistication type I (labelled as such in the LCA analyser) as its method of analysis is quite similar to VS2 or the second type of verb sophistication. Both verb sophistication indices of types I and II showed between-group differences in the pilot study, but the second type showed larger effect sizes. In the VS2 section, I explained what the differences between the first and second types of verb sophistication in LCA are and the reason for including the second type. SVV1 (Squared VV1, measured as T^2_{verb} / N_{verb}) is also excluded from the set of measures in this study as it is very similar to VV1 and showed smaller effect sizes than VV1 in the pilot study. ADVV or the Adverb Variation and MODV or the Modifier Variation measures were also excluded from the analysis as both showed trivial effect sizes (close to zero) in the pilot study. The adverb variation measure also did not show any between-proficiency differences in Paquot's (2019) analysis of academic research papers. The simple count of types as a measure is also excluded from the final set of measures. Unlike other indices in this study which take the token counts as the basis for the ratio or word segments, the simple number of types does not have a basis and therefore, comparing it across rhetorical sections with varying text lengths (or token counts) and even including it in statistical modelling alongside other measures would have been problematic. Table 5.3 shows the selected set of 22 lexical complexity measures in this study.

Table 5.3. The set of 22 lexical complexity measures used in this study

Lexical Measure Label	Attributes	Quantification Method
LD	Lexical Density (Ure, 1971; Engber, 1995)	N_{lex} / N
LS1	Lexical Sophistication type I (Linnarud, 1986)	N_{slex} / N_{lex}
LS2	Lexical Sophistication type II (Laufer, 1994)	T_s / T
VS2	Verb Sophistication type II (Chaudron & Parker, 1990)	T^2_{sverb} / N_{verb}
CVS1	Corrected Verb Sophistication type I (Wolfe-Quintero et al., 1998)	T_{sverb} / $\sqrt{2N_{verb}}$
NDWERZ	Number of Different Words (Malvern et al., 2004)	Means of NDW for 10 random sub-samples of 50 words

Lexical Measure Label	Attributes	Quantification Method
NDWESZ	Number of Different Words (Malvern et al., 2004)	Means of NDW for 10 random sub-samples of 50 consecutive words with random starting points
MSTTR	Mean Segmental TTR (Johnson, 1944)	Means of TTRs for 50-word segments
MATTR	Moving-Average TTR (Covington & McFall, 2010)	TTRs of fixed-length successive moving windows
MTLD	Measure of Textual Lexical Diversity (McCarthy, 2005; McCarthy & Jarvis, 2010)	Mean length of word strings with TTR of 0.72, words/factors method
LOGTTR	Bilogarithmic TTR, Herdan's C	Log T / Log N
UBER	Uber's U (Dugast, 1978)	$\text{Log}^2\text{N}/(\text{LogN} – \text{LogT})$
Maas	Logarithmic corrections of a^2, lgV0, and lgeV0 (Maas, 1972, cited in Treffers-Daller, 2013)	$(\text{Log N} – \text{Log T}) / \text{Log}^{2}\text{N}$
Vocd-D	The adapted D (Malvern & Richards, 1997; Malvern et al., 2004)	Random sampling of words for TTR segments, curve-fitting method
HD-D	Hypergeometric D (McCarthy & Jarvis, 2007)	Sum of lexical probabilities based on random samples of 42 words
RTTR	Root TTR, Index of Guiraud (Guiraud, 1954)	$T / \sqrt{N}$
LV	Lexical Variation (Linnarud,1986)	T_{lex} / N_{lex}
NV	Noun Variation (McClure, 1991)	T_{noun} / N_{lex}
ADJV	Adjective Variation (McClure, 1991)	T_{adj} / N_{lex}
VV1	Verb Variation type I (Harley & King, 1989)	T_{verb} / N_{verb}
CVV1	Corrected Verb Variation I (Wolfe-Quintero et al., 1998, as an adaptation of Carroll's (1964) CTTR method	$T_{verb} / \sqrt{2}N_{verb}$

Lexical Measure Label	Attributes	Quantification Method
VV2	Verb Variation type II (McClure, 1991)	T_{verb} / N_{lex}

5.3.1.2. Syntactic Complexity Measures and their Quantification Methods

This study adopts a set of 11 syntactic complexity measures with 4 broad categories or constructs of 'length of production unit', 'amount of subordination', 'amount of coordination', and 'degree of phrasal sophistication' as presented in the following paragraphs. These measures gauge the clausal, phrasal, and T-unit-based aspects of the syntactic structures that have been strongly recommended to cover all dimensions/aspects of global-level syntactic complexity in SLA research (e.g., Lu, 2017; Norris & Ortega, 2009). All the measures that are selected for this study and described in the following paragraphs are tested for reliability (both between-annotator agreement and between annotators and the L2SCA) by Lu (2010); a similar reliability test has also been conducted by Yoon and Polio (2016). Reliability and validity of these syntactic measures have been further confirmed by Polio and Yoon (2018), especially regarding genre differentiation (see section 3.3). Therefore, I will not repeat the evidence of their reliability and validity hereafter. Lu (2017) reports a number of studies that show the syntactic measures in L2SCA are predictive of holistic measures of writing quality. The statistical results of the pilot study on syntactic analysis of dissertation abstracts will be provided in chapter six.

Length of Production Unit:

The two syntactic measures of mean length of clause and mean length of T-unit represent the construct of length of production unit in this study. Detailed descriptions of these measures and constructs along with a review of studies that used these indices or reported their validity have already been presented in chapter two, section 2.3.1 and chapter three, section 3.3 and 3.4.

Mean Length of Clause (MLC) measures the average number of words in each clause. Several studies reported a relationship between the increased length of clauses and higher proficiency levels (e.g., Ai & Lu, 2013; Kim, 2014; Kyle, 2016; Liu & Li, 2016; Lu, 2010; Ortega, 2003; Wolfe-Quintero et al., 1998; Yoon, 2017). Both Ai and Lu (2013) and Lu and Ai (2015) also found statistically significant differences between the combined non-native speakers' groups (EFL with various L1s) and English native speaker group in the MLC values.

Mean Length of T-unit (MLT) is another length-based measure which counts the length of T-units and was first proposed by Hunt (1965) as a main clause together with other dependent clause/non-clausal structures that are attached to it. Ortega (2003) and Wolfe-Quintero et al. (1998) are among the notable studies which reported positive relationships between writing proficiency and the increased length of T-units. Likewise, both Lu (2010, 2011) and Yoon (2017) confirmed significant between proficiency-level differences in the production of longer T-units in argumentative essays, Ai and Lu (2013) show significant differences between the combined as well as individual NNS (EFL) groups and the NS group in academic essays, Kim (2014) documented statistically meaningful differences across the three basic, intermediate and advanced proficiency levels in academic argumentative essays, Mancilla et al. (2015) reported a linear increase in the mean values of this measure in the texts of low and high proficiency ESL graduate students and the NS group (English L1s), and finally, Yang et al. (2015) found this index to be a good indicator of ESL writing quality judged by human raters. In the pilot study also this measure showed a significant difference between EFL and English L1 dissertation abstracts as confirmed via three separate post-hoc comparison tests.

Kyle (2016) reported a strong correlation (r = 0.8) between MLC and MLT indices. This is while MLC gauges sub-clausal complexity and is affected especially by an increase in phrasal coordination (Kyle, 2016) but MLT more specifically gauges complexity by subordination (Bardovi-Harlig, 1992; Ortega, 2000). Nontheless, Ortega (2003) confirmed that both indices are reliable indicators of L2 writing proficiency.

Amount of Subordination

The four indices of clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, and dependent clauses per T-unit represent syntactic subordination in this study. A detailed explanation of these indices as well as an in-depth review of the studies that employed them or reported their validity have been presented in sections 2.3.2, 3.3., and 3.4.

Clauses per T-unit (C/T) index (also labelled as C.T in the tables and graphs in this thesis) calculates the number of clauses in each T-unit. This includes all types of clauses and thus does not differentiate between types of subordination. Six of the studies (of the total 18 studies) reviewed by the Wolfe-Quintero et al. (1998) also reported a significant and positive relationship between the C/T values and language proficiency. However, Lu (2011) and

Knoch et al. (2014, cited in Kyle, 2016) did not find meaningful relationships between this measure’s values and language proficiency and development. Lu and Ai (2015) also did not find any difference in the C/T values between EFL and English L1 college students’ argumentative essays. An earlier study, Flahive and Snow (1980), however, is among the studies that showed this measure can discriminate between proficiency levels and that it has a positive relationship with the quality of ESL texts (evaluated based on holistic ratings). Ortega (2003), likewise, concluded that this index is a reliable indicator of proficiency-level differences in L2 writing. The results of the pilot study also point to the effectiveness of this measure in capturing between-group differences in academic writing; EFL group produced significantly lower amount of clauses per T-unit than English L1.

Complex T-units per T-unit (CT/T) (also labelled as CT.T in the tables and graphs in this thesis) measures the number of complex T-units in each T-unit. A complex T-unit should at least have one dependent clause. The CT/T ratio, therefore, counts the number of T-units that have dependent clauses but does not differentiate between the types of these dependent clauses nor does it consider how many dependent clauses are present in that T-unit. Casanave (1994) found a positive trend between this measure’s values and language development. Kim (2014) also found a linear positive increase (with a large effect size) in the use of complex T-units in total T-units among the three proficiency levels of college-level EFL students’ writing. Her results also show that this index was a strong predictor of L2 English academic writing proficiency. Similar results for this measure are obtained by Lu and Ai (2015) regarding the differences between English L1 and L2 groups’ performances. CT/T is also one of the syntactic measures which showed differences between the EFL and English L1 groups’ academic writing in the pilot study.

Dependent Clauses per Clause (DC/C) (also labelled as DC.C in the tables and graphs in this thesis) which measures the number of dependent clauses per clause is another syntactic measure indicating clausal subordination. Mancilla et al. (2015) reported a significant difference between the English L1 and ESL graduate students’ texts regarding the values of this measure. This result is consistent with the Ai and Lu’s (2013) findings of academic writing where the English L1 group showed a significantly higher mean value than both EFL groups at low and high proficiency levels, as well as with Kim’s (2014) study of college-level writing where this measure’s values were significantly different across the three EFL proficiency levels. Lu (2011) on the other hand, showed a mixed result of subordination for the EFL students’ argumentative writing where the values increase during the first two years and decrease over the last two years of university. In a follow-up study, Lu and Ai

(2015), on the contrary, showed that English L2 groups with different L1s produced significantly less dependent clauses (DC/C) than English L1s. In the specialised academic writing corpus of the pilot study also English L1 group used a significantly larger proportion of dependent clauses in total clauses than the EFL group.

Dependent Clauses per T-unit (DC/T) (also labelled as DC.T in the tables and graphs in this thesis) measure is similar to the above measure of DC/C in that both calculate the number of dependent clauses, so it is assumed that both measures are highly correlated. This is in fact confirmed by a number of studies where both measures showed either significant differences or no difference between proficiency levels or between English L1 and L2 groups (Ai & Lu, 2013; Kim, 2014; Lu, 2010, 2011; Lu & Ai, 2015; Mancilla et al., 2015; Nasseri, 2017). In Lu (2010) these two measures were correlated at r = 0.92. Lu (2011) found a negative relationship between the obtained values of DC/T and language proficiency while Hamburg (1984, cited in Kyle 2016) and Kim (2014) showed a positive relationship between these two variables. Moreover, Ai and Lu (2013), Mancilla et al. (2015) as well as the pilot study of this research all found meaningful differences between the English L1 and EFL/ESL groups' production of this measure in academic writing. I will examine which of these two measures could capture between-group differences in this study's various rhetorical sections and whether any of the subordination measures could be strong predictors of rhetroical section and group membership in chapter six, section 6.2.6.

Amount of Coordination

The two indices of coordinate phrases per clause and coordinate phrases per T-unit are selected to represent the construct of syntactic coordination in this study. Full accounts of these and other coordination structures and indices have been presented in chapter two, section 2.3.3. Evidence for the usefulness of these indices has also been presented in 3.4.

Coordinate Phrases per Clause (CP/C) (also labelled as CP.C in the tables and graphs in this thesis) is a syntactic measure of phrasal coordination which calculates the number of coordinate phrases to the total number of clauses. Based on the specification of Cooper (1976), coordinate phrases only include noun, verb, adjective and adverb phrases (those that immediately dominate a coordinating conjunction). Kyle (2016) noticed when the clause length increases, students who used more coordinate phrases received higher scores. This measure has produced mixed results in different studies and, therefore, is included in this study's long dissertations texts to examine its efficacy as an indicator of syntactic proficiency. Lu (2010) is among the studies which showed differences across the three proficiency levels in academic writing among English major students. Mancilla et al. (2015) study of graduate students' texts showed that the ESL group outperformed the English L1 group with regard to the production of this measure. Other studies (e.g., Ai & Lu, 2013; Kim, 2014; Lu & Ai, 2015), however, did not report any between-group or between proficiency-level differences regarding the production of coordinate phrases per clause.

Coordinate Phrases per T-unit (CP/T) (also labelled as CP.T in the tables and graphs in this thesis) is another phrasal coordination measure which captures the ratio of coordinate phrases to the number of T-units. Just as CP/C, it includes all types of coordinate phrases and does not capture individual types. These two measures are highly correlated (r = 0.94, Lu, 2010). Kim (2014) and Lu (2010) both reported this measure as an indicator of L2 writing proficiency across three proficiency levels. Likewise, findings of Ai and Lu (2013) demonstrate this index among the measures which distinguish between English L1 and EFL university students' writing (results for the combined as well as separate EFL groups). The rather short texts of abstracts of MA dissertations in the pilot study did not reveal any statistically significant differences between the three postgraduate groups.

Even though Lu (2017) stated that larger values of all syntactic complexity indices in L2SCA including coordination indices denote higher degrees of syntactic complexity, I have put forth in sections 2.3.3 and 3.2 the collective evidence in the related scholarship that larger amounts of coordination structures are usually associated with lower levels of linguistic proficiency. Regarding the exact measures of CP/C and CP/T, there are mixed results concerning the association of coordination with proficiency (see for instance the conflicting results in Ai & Lu, 2013, Lu, 2011, and Lu & Ai, 2015). Since I did not have access to a formal record of linguistic proficiency of the students who wrote the dissertations, in this study I will investigate this matter from (postgraduate) English L1 vs. L2 point of view. Ortega (2000) has also documented a decrease in coordination with an increase in subordination. I will examine if similar patterns can be seen in the results of predictive statistical modelling across groups and rhetorical sections in chapter six. I will further examine if coordination is a distinct feature of any of the six rhetorical sections of MA dissertations.

Degree of Phrasal Sophistication

The three measures of complex nominals per clause, complex nominals per T-unit, and verb phrases per T-unit are selected to gauge phrasal complexity of the three groups' dissertations. Phrasal-level structures have been specifically regarded as distinct aspects of advanced-level and specialised academic writing and as such, phrasal complexity indices are considered as reliable indicators of proficient writing and as predictors of academic writing quality (Biber & Gray, 2013, 2016; Biber, Gray, & Poonpon, 2011; Bulté and Housen 2014; Gray, 2015; Liu & Li, 2016; Yoon, 2017). More details about phrasal complexity and/or sophistication, and other phrasal structures have been presented in chapter two, section 2.3.4.

Complex Nominals per Clause (CN/C) calculates the ratio of the number of complex nominals to the number of clauses. CN/C demonstrated statistically significant differences between the three proficiency levels in Lu (2010) and Kim (2014), between the NS and EFL groups in Lu and Ai (2015), and between the combined and individual NNS groups and the NS group in Ai and Lu (2013). Yoon (2017) study is another example of recommending this index as a predictor of L2 writing proficiency. Yoon (2017) also noticed an increase in the number of complex nominals in the academic writings of Chinese EFL learners compared to clausal embeddings as proficiency level increases. Interestingly, the ESL group in Mancilla et al. (2015) study of graduate students' texts outperformed the NS group in the production of complex nominals per clauses.

Complex Nominals per T-unit (CN/T) is a similar measure to the CN/C measure but calculates the ratio of the number of complex nominals to the number of T-units. The findings of Kim (2014), Ai and Lu (2013), Lu (2010), Lu and Ai (2015), as well as the results of the pilot study, all show significant between-group and/or between proficiency-level differences regarding the values of CN/T. Liu and Li (2016) also found that MA students produced fewer complex nominals per T-unit in their dissertations than writers of published research articles.

Both measure capture the following three categories as complex nominals based on the specification of Cooper (1976): 1- nominal clauses, 2- nouns plus an adjective, participle, appositive, possessive, prepositional phrase, and relative clause, and 3- gerunds and infinitives in the subject position. Kyle (2016) noticed as the length of T-unit and clauses increase, students use more complex nominals. Both measures are highly correlated (*r* is above 0.8) as shown in Lu (2011) and Kyle (2016). Among the two measures, it seems that CN/C performs better at capturing proficiency differences as is shown and recommended by Lu (2011) and Wolfe-Quintero et al. (1998). Other scholars such as Liu and Li (2016) and Kim (2014) suggest that both CN/C and CN/T are strong indicators of proficiency and development in syntactic complexity in academic writing. I will examine which of these two indices could better capture syntactic proficiency differences of the three groups' academic writing, and which index among the two could be a better predictor of group and rhetorical section membership based on predictive models as discussed in 6.2.6.

Verb Phrases per T-unit (VP/T) index which measures the number of verb phrases (both finite and non-finite verb phrases) in T-units was proposed by Wolfe-Quintero et al. (1998). It is shown to distinguish between the English L1 and EFL groups' academic writing in the pilot study, and across three proficiency levels in Kim (2014) study of English L2 undergraduate students' writing. This measure has not been extensively used in SLA or writing research studies. It is not clear if this index has high correlations with other verb-based measures of lexical variation (e.g., CVV1, VV1, and VV2) and the two verb sophistication measures (VS2 and CVS1) and that whether the number of verb phrases significantly differ in various rhetorical sections of dissertations produced by the students with different English language backgrounds. Table 5.4 presents the set of 11 syntactic complexity measures used in this study.

Table 5.4. The set of 11 syntactic complexity measures investigated in this study

Syntactic Measures' Labels	**Attributes**	**Quantification Method**
MLC	Mean Length of Clause (e.g., in Lu, 2010; Wolfe-Quintero et al., 1998)	Number of words/ Number of clauses
MLT	Mean Length of T-unit (e.g., in Hunt, 1965)	Number of words/ Number of T-units
C/T	Clauses per T-unit (e.g., in Lu, 2010; Wolfe-Quintero et al., 1998)	Number of clauses/ Number of T-units
CT/T	Complex T-units per T-unit (e.g., in Casanave, 1994; Lu, 2010, 2011)	Number of complex T-units/ Number of T-units
DC/C	Dependent Clauses per Clause (e.g., in Kyle, 2016; Lu, 2010, 2011; Mancilla et al., 2015)	Number of dependent clauses/ Number of clauses
DC/T	Dependent Clauses per T-unit (e.g., in Ai & Lu, 2013; Alexopoulou et al., 2017)	Number of dependent clauses/Number of T - units
CP/C	Coordinate Phrases per Clause (e.g., in Lu, 2010, 2011; Mancilla et al., 2015)	Number of coordinate phrases/ Number of clauses
CP/T	Coordinate Phrases per T-unit (e.g., in Ai & Lu, 2013; Kim, 2014)	Number of coordinate phrases/ Number of T-units
CN/C	Complex Nominals per Clause (e.g., in Kyle,	Number of complex

	2016; Liu & Li, 2016; Lu, 2010)	nominals/ Number of clauses
CN/T	Complex Nominals per T-unit (e.g., in Ai & Lu, 2013; Liu & Li, 2016)	Number of complex nominals/ Number of T-units
VP/T	Verb Phrases per T-unit (e.g, in Kim, 2014; Lu, 2011;Wolfe-Quintero et al., 1998)	Number of verb phrases/ Number of T-units

Among the set of syntactic measures provided in the L2SCA analyser (see 5.3.2), a few measures are left out from the final set of syntactic measures in this study. MLS or the Mean Length of Sentence is not included in the measure set as it produced identical results with the MLT measure of the length of T-units in the pilot study of dissertation abstracts and is very highly correlated with MLT (r = 0.90) in Lu (2010). The other reason that I dropped this measure out was that no other syntactic measure in this study has the ‘sentence (S)’ as the denominator or the base of measurement; therefore, the interpretation of results of this measure would not have been as meaningful as MLT since seven measures in this study incorporate T-unit in their formulas. Other issues with the MLS measure are the probability of the presence of multiple T-units in a sentence as well as the presence of run-on sentences which affect the MLS counts. The T/S measure was also dropped from the L2SCA set of syntactic measures as no study so far reported any between-group or between proficiency-level differences regarding the production of this measure, nor has any study, to my knowledge, confirmed this measure as an indicator of English language development and proficiency. Only one study (Monroe, 1975) confirmed this measure as an indicator of language proficiency in his research on syntactic proficiency in French. The next measure that I did not include is the C/S measure which calculates the number of clauses per sentences. This measure also did not show (except in Kim, 2014 study with a not large enough effect size) to be a good indicator of language proficiency and development. Some of the syntactic measures that were reviewed in chapter two were not included in the final set of indices either because of insufficient evidence as to whether they could discriminate between groups with various English language backgrounds in advanced and/or academic English texts, or because at the time of the final analysis of this study, I did not have any access to the programmes that could automatically compute them.

5.3.2. The Programmes Used to Analyse the Measures

To maintain the criteria for word count and calculating the indices, and to ensure the validity of the comparison of findings of indices across platforms, care has been taken to conform the tokenisation, tagging and lemmatisation processes across the analysers as listed below:

- The PTB English Tokenizer (Manning et al., accessed 2018) or PTB tokenisation style (style based on the Penn Treebank project): punctuations are not considered as words even though they are assigned separate tags; contracted forms and possessive forms are considered as separated tokens. The PTB tokeniser is deterministic but it has some good heuristics to deal with single quotes as part of the words, periods as part of the words vs. as end-of-the-sentence marker, etc.
- Taggers with the left3words tagging model trained on the Penn Treebank Tagset (Marcus, Santorini, & Marcinkiewicz, 1993) based on the WSJ (Wall Street Journal Corpus).
- Lemmatisation with Morpha (Minnen, Carroll, & Pearce, 2001). This is a morphological analyser for English which includes a verb-stem list of verbs that have doubling of consonants (for example for British English). Morpha takes as input the already-POS-tagged files.

Four NLP analysers and programmes are used to compute the lexical and syntactic complexity measures in tables 5.3 and 5.4, as listed below:

TAALED (Tool for the Automatic Analysis of Lexical Diversity; Python-based beta version 1.2.4, Kyle, 2018) was used to calculate the measures of MSTTR, MATTR, HD-D, MTLD, and MAAS. This version of TAALED uses Stanford CoreNLP (version 3.5.1) using the Maxent Tagger with the above-mentioned tagging specifications as well as the morpha class for stemming and morphological processing. All indices are calculated using the lemma forms.

Coh-Metrix (version 3.0; Graesser, McNamara, Louwerse, and Cai, 2004) text analysis tool was used to obtain the original D values, i.e, vocd-D index of lexical diversity. The original tool which was shared by the developers privately takes an entire corpus with text files as the input and outputs a .csv file with the Vocd-D values in a separate column. McCarthy and Jarvis (2010) validated the lexical diversity indices in *Coh-Metrix*. For POS tagging, CohMetrix uses the Charniak parser (Charniak & Johnson, 2005) which is a part of the Stanford NLP Parser based on the Penn Treebank annotation guidelines. The vocd-D is calculated based on the word forms. To conform this with other measures in LCA-AW (see below) and TAALED which are calculated using lemma forms, the lemmatised files (.lem files as lemmatised by Morpha) can be given as input to CohMetrix rather than the text files.

The Lexical Complexity Analyzer for Academic Writing (***LCA-AW*** version 2.1; Nasseri & Lu, 2019) was used to analyse lexical density (LD), NDWERZ, NDWESZ, Herdan's C or LOGTTR, Uber's U, Guiraud's R or RTTR, LV, NV, ADJV, VV1, CVV1, and VV2 measures of lexical diversity, and LS, VS2 and CVS1 measures of lexical sophistication. *LCA-AW* is a modified version of the LCA (Lexical Complexity Analyzer, Lu, 2012). LCA was developed and reliability tested by Lu (2012, 2014). Both web-based and downloadable versions of the LCA analyser takes the BNC (the British National Corpus) or ANC (American National Corpus) and their respective frequency word lists, which are general English word lists, as the reference point to calculate lexical sophistication indices as the ones which do not appear in the first 2000 most-frequently-used words in the BNC or ANC word lists. Since my study analyses a discipline-specific academic writing corpus, I included the BAWE (British Academic Written English) corpus and its most-frequently-used academic writing words used in linguistics and language studies as well (see sections 2.2.3 and 7.2.2). More details about the LCA-AW programme and the way to download and use it will be presented in Appendix D.

The new version, LCA-AW, integrates the BAWE word list along with the BNC (with an option to change to the ANC) as filters for calculating lexical sophistication indices. It takes the academic writing corpus texts and calculates the sophisticated words as the ones that neither appear in the top 2000 frequently-used BNC word list nor in the top 100 frequently-used academic writing word list for linguistics-related disciplines. The entire corpus can be processed via the *folder-lc.py* script. The analyser requires the pre-processing of files separately for POS tagging and lemmatisation. I used Stanford POS Tagger (version 2015. 01. 30; Toutanova et al., 2003) for tagging the files and Morpha (Minnen et al., 2001) to lemmatise them.

The frequency of types and tokens of eight lexical units is obtained and based on that the above-mentioned lexical complexity measures are computed. Regarding the measurement criteria, LCA-AW has the following specifications:

- All indices are calculated using the lemma forms,
- Punctuations are not counted as tokens even though they receive separate tags from the tagger,

- Different inflections of a lemma are counted as one type,
- Lexical words are specified as nouns, adjectives, and verbs (except modals and auxiliary verbs of 'be' and 'have'), and lexical adverbs which in the BNC word list are both adjective and adverbs, as well as adverbs with adjectival roots and adverbs with -ly suffixes.
- LS2 uses all sophisticated types (i.e., unique words) but LS1 uses only lexical (i.e., content words) tokens and sophisticated lexical tokens.

The Syntactic Complexity Analyzer (***L2SCA,*** version 2014-01-04; Lu 2010) was used to analyse 11 measures of syntactic complexity as discussed in section 5.3.1.2. The system takes plain texts or an entire folder containing text files as input and pre-processes the files first via the Stanford Parser (Klein & Manning, 2003). This syntactic parser has in-built sentence segmentation (to separate sentences each on a new line), tokenisation (to separate each token, e.g., words, punctuation marks, acronyms, and numbers), and POS tagging (to assign part-of-speech categories to each word, e.g., noun, adjective, etc) functionalities to syntactically parse the texts to produce parse trees. The analyser then counts the frequencies of the following nine basic production units and syntactic structures in each text: words, sentences (S), clauses (C), dependent clauses (DC), T-units (T), complex T-units (CT), coordinate phrases (CP), complex nominals (CN), and verb phrases (VP).

Words are counted as tokens which are not punctuation marks. The other eight units are counted via Tregex (Levy & Andrew, 2006) by querying the parse trees (for the Tregex patterns and the definitions of the nine production units see Lu, 2010). Based on the frequency counts of all these syntactic units, L2SCA calculates the syntactic measures outlined in the previous section as the ratio of one syntactic unit to another and outputs the files with lines of comma-delimited lists of values of the measures. Lu (2010) reported a high degree of reliability for production units (using two inter-annotators' values against the values obtained from the system) with F-scores of 0.84 for complex nominals and 1 for sentences on the development data. A high degree of reliability is also achieved for the values/scores of syntactic measures, with a correlation of 0.84 for CP/C and 1 for MLS on the development data. Lu (ibid.), further confirms that learner writing errors (e.g., issues with agreement or determiners) do not result in structural misanalysis by the parser or misrecognition of the syntactic units by the system.

The values that were obtained from these four programmes were then subject to a series of statistical tests that will be discussed in the next chapter. As a visual aid, the lexical measures in the text, tables and graphs will be presented in lower-case letters and the syntactic measures in upper-case letters in the following chapters.

6 Statistical Procedures, Results, and Discussions of the Findings

6.1. The Measure-Testing Process and an Overview of this Chapter

In this chapter, I examine the performances of the three groups regarding the production of lexically and syntactically complex texts based on the values of the 32 measures described in the previous chapter. To do this, I first present the descriptive statistics together with a visual inspection of the data to get a grasp of the distribution of the observed values as well as residuals of the measures as will be explained.

In the next step, I present the between-group differences of lexically and syntactically complex texts using analyses of variance and post-hoc comparison tests. The results of these tests will first be demonstrated for the pilot studies and the entire corpus and then in each of the six rhetorical sections for possible patterns regarding the similarities and differences of the texts of the groups in terms of noticeable lexical and syntactic constructs and measures. In this stage, I also examine which of these complexity measures can consistently capture complexity differences of the texts of the three groups. Linguistic examples as excerpts from the dissertations of the three groups will then be qualitatively analysed and compared with the quantitative findings. This is for further insight into the form-function relationships, i.e., the types of lexical and syntactic features produced by the students and the rhetorical functions of those sentences/texts.

This step will be followed by an examination of the relationship between these measures to find if the effective measures for capturing between-group differences are highly correlated. To complement this examination of the relationships between and among these complexity measures, I will carry out a series of factor analyses, both to examine their structures relative to the existing classification of these measures and constructs in the literature, as well as further exploratory analyses.

I will then investigate the effect of a main text-intrinsic variable (rhetorical sections as sub-genres of the texts) and a main text-extrinsic variable (groups of students with different English language backgrounds) on the variation of the values of these selected complexity measures. The final statistical analysis will be dedicated to building predictive models to find which lexical and syntactic measures can better predict/distinguish/classify the groups of students and the rhetorical section/function of the texts of these dissertations.

The objectives of the measure-testing process throughout this chapter are to investigate similarly-calculated measures as well as different measures that represent lexical and syntactic constructs to arrive at the final set of unique lexical and syntactic measures to answer the research questions as well as to verify the performance of the variables based on the previous research. The findings of this step can help subsequent researchers in the selection of a suitable set of measures for similar research designs. The best-performing measures for each of the following statistics are discussed in the answers to research questions in 6.8 and a brief conclusion of the recommended measures is presented in chapter seven. The statistical procedures in the following sections are all carried out using the R programming language (versions 3.5.3 and 3.6.0; R Core Team, 2013). Supplementary data and results are provided in Appendix B, including a link to the R code.

Prior to carrying out the analyses related to measure-testing, the data points corresponding to the six sub-genres (rhetorical sections) from every student had to be aggregated to one data point. This is to comply with the assumption of independence of data points (see, for example, Winter, 2019) which is a prerequisite for performing any inferential statistics. Since various sub-genres of the dissertations had different token counts (and hence different weights or values), I aggregated the data based on the weighted mean method, taking the token count as the weight. For this step, the *data.table* package (version 1.12.0, Dowle, 2019) and the *stats* package (version 3.6.0, base R) were used. The two versions of genre-aggregated and genre-separated datasets were prepared.

The following paragraphs explain in detail the statistical procedures and tests used to answer the research questions specified in 1.4.5. Each section begins with an explanation of the relevant statistical test and a description of that test, including the details of statistical procedures and the results. At the end of each section I present in-depth discussions of the findings in light of previous related works. The results of multiple tests collectively will be used to answer/interpret the research questions from different angles in section 6.8.

6.2. Descriptive Statistics

Many statisticians and applied linguists strongly recommend a visual inspection of the data for assessing the normality of the distributions and the homogeneity of variances instead of conducting formal tests (for detailed discussions against conducting such formal tests see Larson-Hall, 2016; Zuur, leno, & Elphick, 2010; Wilcox, 2011; Winter, 2019). To compensate for the presence of skewed data and outliers, I used robust statistics such as non-parametric bootstrapping methods that do not depend on the assumptions of normality of the data and/or the homogeneity of variances.

To get a grasp of the distributional properties of lexical and syntactic measures, descriptive statistics were obtained. This step was accompanied by creating histograms for visual diagnostics. Tables 6.1 and 6.2 show the main descriptive statistics for lexical and syntactic measures respectively. Descriptive statistics based on individual groups in each rhetorical section will be provided in the monofactorial tables throughout 6.3. Graphs 6.1. and 6.3 also demonstrate the distribution of the data in the lexical and syntactic datasets respectively.

The descriptive statistics illustrated in these graphs and tables indicate that most of the lexical and syntactic measures have somewhat normal distributions, with maas, mattr, msttr, logttr, and nv having more normally-distributed values (e.g., closer to Gaussian distribution) among the lexical indices, and CP/C (Coordinate Phrases per Clause) and CP/T (Coordinate phrases per T-unit) being more normally-distributed among the syntactic indices. Adjv (adjective variation), ld (lexical density), vocd (a variant of the D-measure), and mtld (measure of textual lexical diversity) measures are among the heavy-tailed lexical indices, and MLT (Mean Length of T-unit), DC/T (Dependent Clauses per T-unit), VP/T (Verb Phrases per T-unit), C/T (Clauses per T-unit), and CN/T (Complex Nominals per T-unit) are the syntactic heavy-tailed ones. However, the assumption of normality (e.g., normal or skewed distributions for linear models like ANOVA and mixed-effects models as will be discussed in this chapter) is not to be met for the data itself but for the residuals. That is, a measure's datapoints could have a skewed distribution but a normal distribution of the residuals. The residuals of a measure are the differences between the observed data points and the predicted/ fitted data points as computed via a regression model. These differences (e.g., residuals) need to be (approximately) equally distributed across the predictor variable (e.g., the groups in this case) for the assumption of homoscedasticity to be met. To check the normality of the residuals of the indices, I first obtained regression models for all measures based on the groups as the predictor variable.

The residuals were then extracted from these models and plotted in quantile-quantile (Q-Q) plots as demonstrated in graphs 6.2 and 6.4 for lexical and syntactic measures respectively. A quantile is the percent of data points below a given value. A Q-Q plot of residuals is a plot of the quantiles of the observed data (labelled as 'sample' in the graphs) against the predicted/fitted data (labelled as 'theoretical' in the graphs). A reference line is also plotted as a guide to check the (equal) distribution of these quantiles. The two graphs confirm homoscedasticity for all lexical and syntactic measures and the absence of any significant deviations from the normal distribution of the residuals.

Table 6.1. Descriptive statistics for the genre-aggregated lexical dataset

Measures	**ld**	**ls1**	**ls2**	**vs2**	**cvs1**	**ndwerz**	**ndwesz**	**maas**	**logttr**	**uber**	**rttr**
Mean	0.034	0.44	0.43	1.78	0.86	40.04	38.08	0.05	0.81	19.3	13.1
SD	0.01	0.1	0.04	0.9	0.2	0.9	1.2	0.0	0.01	1.4	1.6
Median	0.031	0.43	0.43	1.60	0.84	40.15	38.16	0.05	0.81	19.2	12.9
Measures	**lv**	**vv1**	**cvv1**	**vv2**	**nv**	**adjv**	**mattr**	**msttr**	**mtld**	**vocd**	**hdd**
Mean	0.36	0.43	1.40	0.09	0.33	0.03	0.73	0.73	53.7	98.6	0.79
SD	0.08	0.1	0.3	0.02	0.08	0.01	0.02	0.02	8.2	14	0.01
Median	0.37	0.44	1.42	0.09	0.33	0.02	0.73	0.73	52.3	97.9	0.79

--The number of observations for all measures is 210. The measures have different scales/metrics.

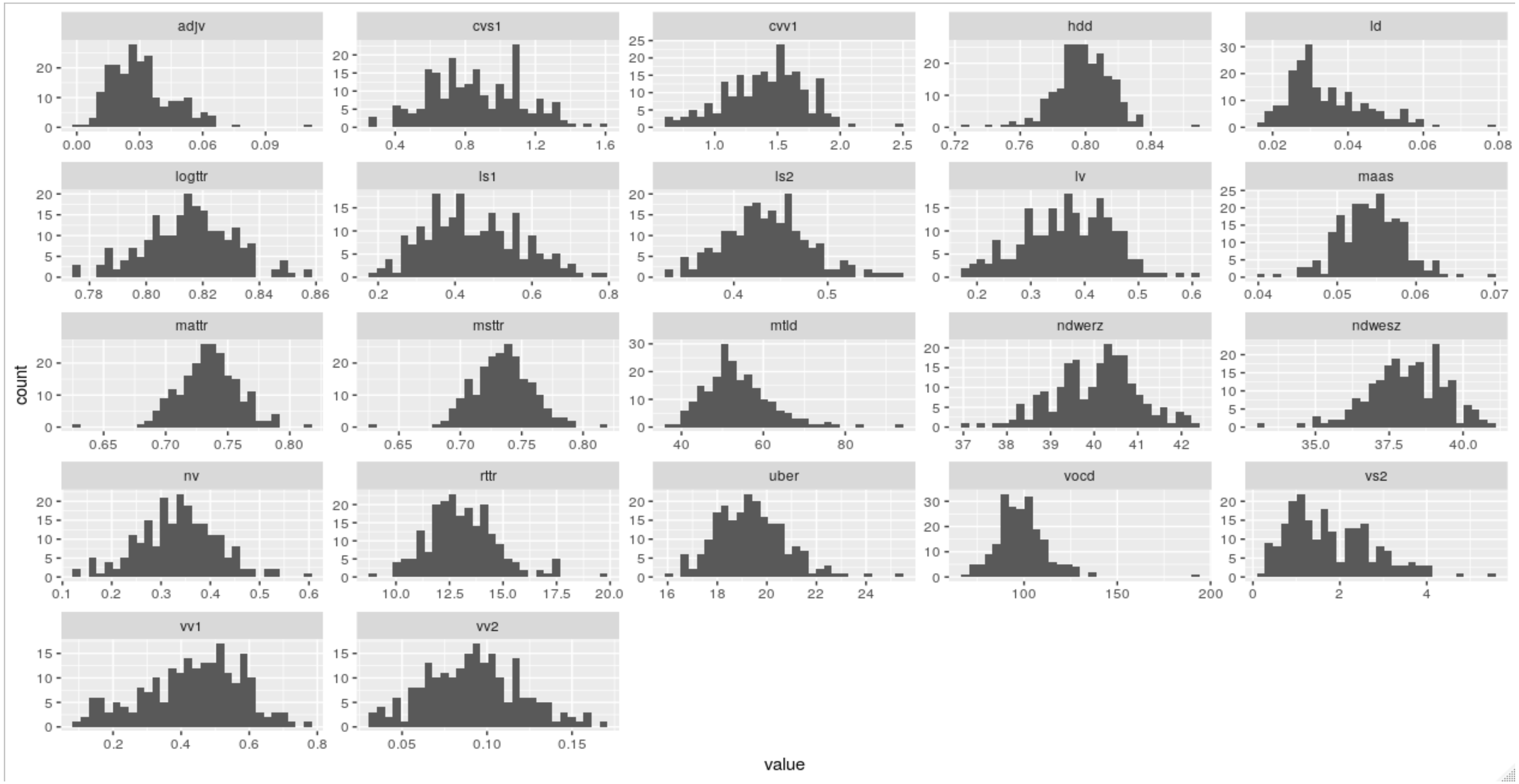
Table 6.2. Descriptive statistics for the genre-aggregated syntactic dataset

Measures	**MLC**	**MLT**	**C.T**	**CT.T**	**DC.C**	**DC.T**	**CP.C**	**CP.T**	**CN.C**	**CN.T**	**VP.T**
Mean	12.29	22.4	1.83	0.48	0.38	0.73	0.38	0.7	1.7	3.1	2.45
SD	1.3	3.3	0.2	0.08	0.05	0.1	0.09	0.1	0.2	0.5	0.3
Median	12.25	21.9	1.79	0.48	0.38	0.68	0.38	0.6	1.7	3	2.40

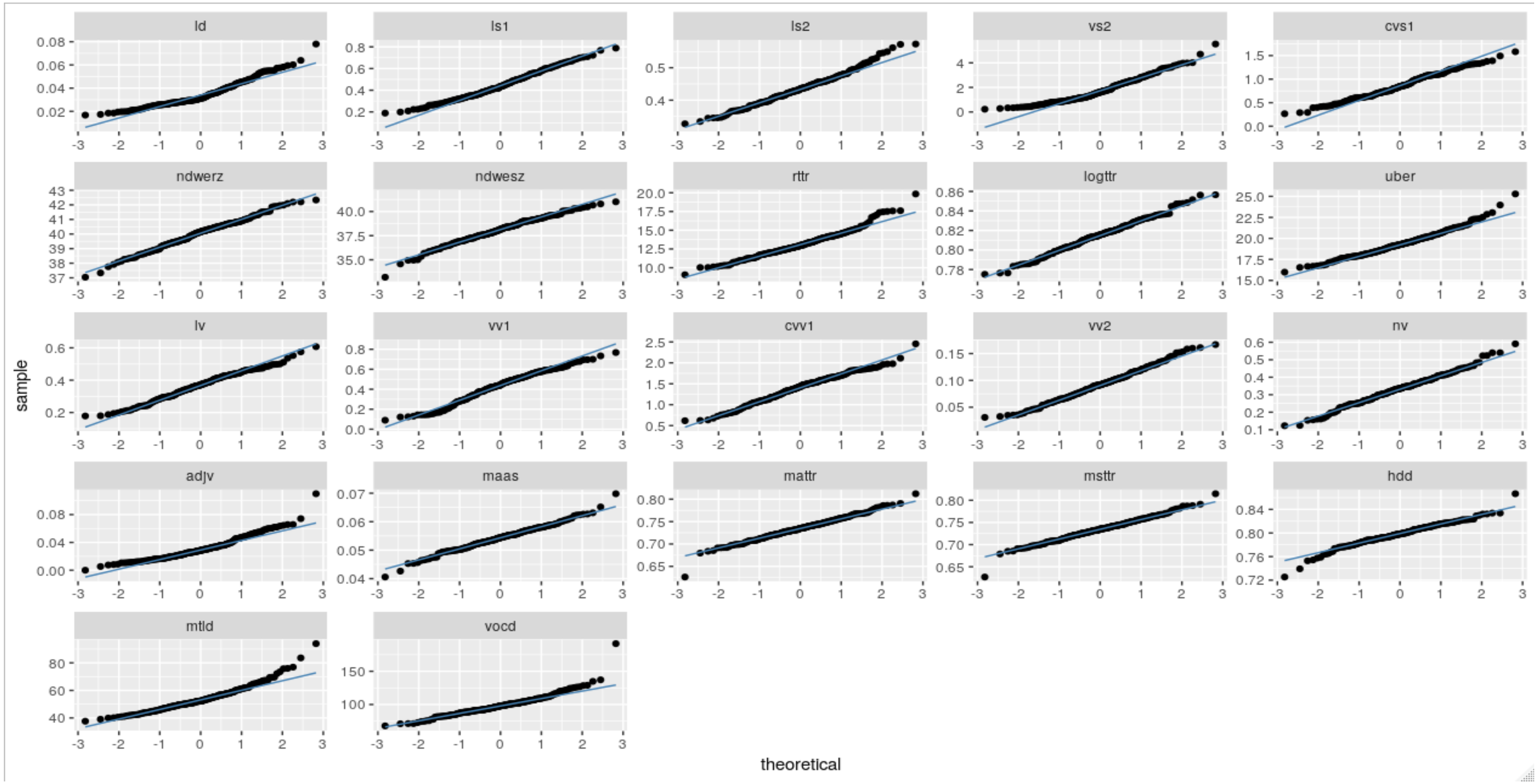
--The number of observations for all measures is 210. The measures have different scales/metrics.

However, as discussed earlier, I use robust statistics such as bootstrapping and the use of bootstrapped confidence intervals and effect sizes for the statistical tests in this chapter to compensate for the presence of outliers, i.e., so that the estimates are not affected by unusually high or low values. This is because the right upper data points in the residuals in the syntactic graphs have some deviations from the line. This indicates, as shown in the histograms, that a handful of texts scored disproportionately higher values in most of the syntactic indices. Upon a manual inspection of the texts, I found that most of these high values belong to the dissertations in TEFL/TESOL and discourse analysis, mainly in the ESL group, followed by the English L1 and EFL groups. Further discussions will be resented regarding the mixed-effects models of the effect of the rhetorical sections on the values of these syntactic indices.

Graph 6.1. Histograms for lexical variables in the entire corpus

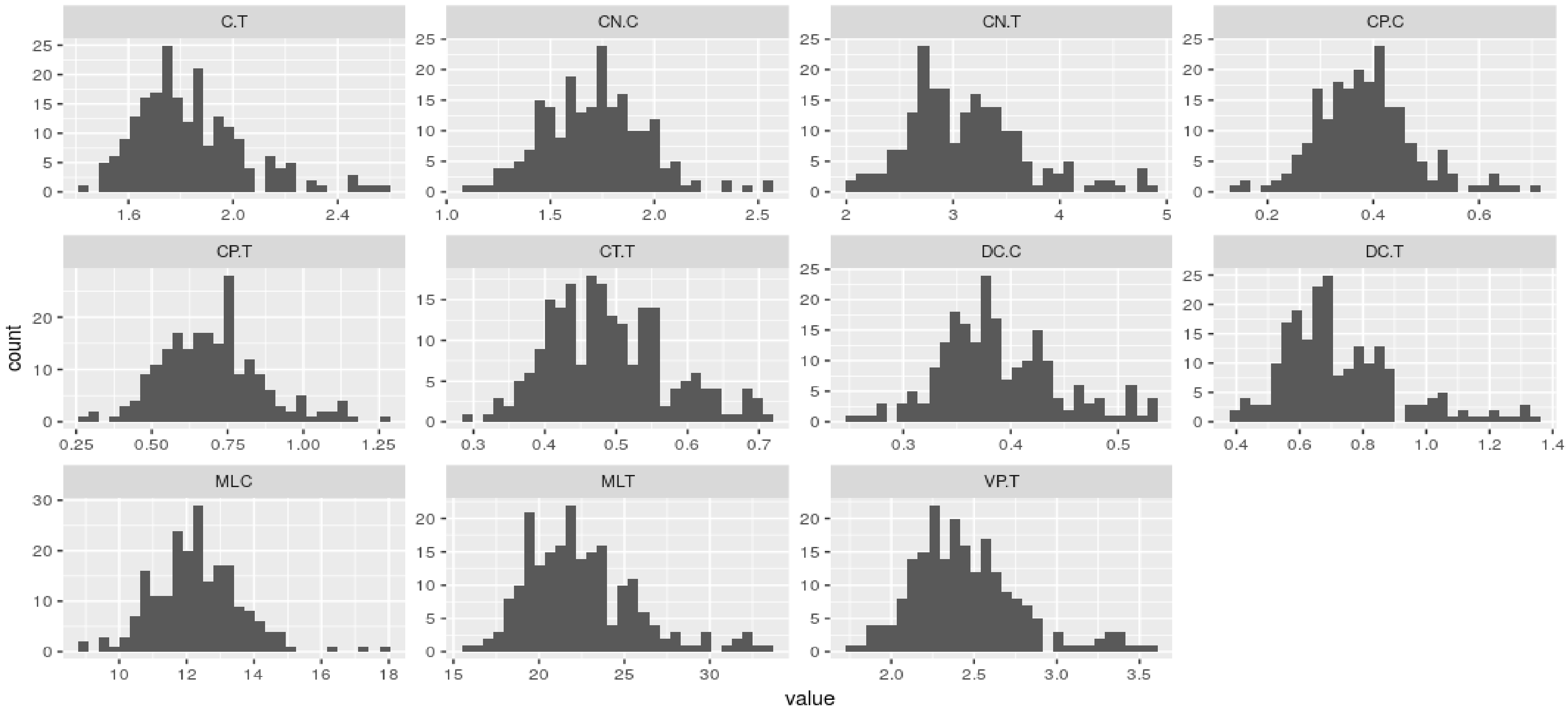


Graph 6.2. Quantile-Quantile plots of residuals of lexical measures

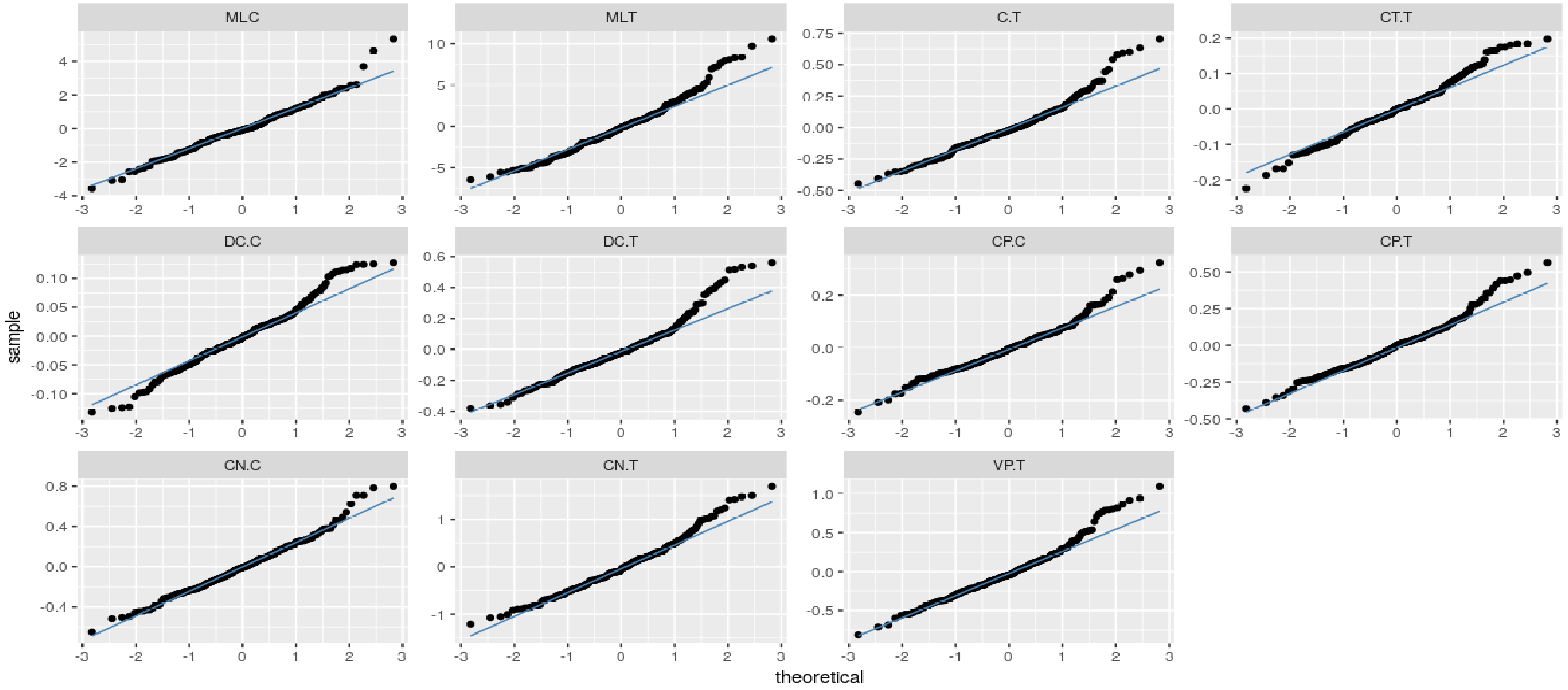


– The residuals are obtained by fitting linear regression models based on ‘groups’ as the predictor variable.

Graph 6.3. Histograms for the syntactic variables in the entire corpus



Graph 6.4. Quantile-Quantile plots of residuals of syntactic measures



– The residuals are obtained by fitting linear regression models based on ‘groups’ as the predictor variable.

6.3. Lexical and Syntactic Complexity Differences of EFL, ESL, and English L1 Groups

To examine whether there are any between-group differences regarding lexically and syntactically complex texts produced by the three groups, a pilot study with a subset of the final data for the abstract sections of the dissertations was conducted. This process was then repeated for the entire corpus (the groups-and genre-aggregated data) to confirm whether any between-group differences exist in the first place. The results of both analyses confirmed between-group differences for at least one set of comparison for several lexical and syntactic measures as will be discussed in detail.

For these tests in the following tables, a series of general linear models (one-way ANOVA or Analysis Of Variance) tests were run to find whether there is any difference between the means of the three groups overall. In the cases where an overall difference was soptted, the ANOVA test is followed by post-hoc multiple comparison tests of Tukey HSD to see any specific significant difference between the pair-wise comparisons. HSD results are accompanied by 95% confidence intervals (CI) of the mean differences using the R code in Laflair, Egbert, and Plonsky (2015, p. 70), as well as the point estimates of Cohen's *d* effects sizes and the confidence intervals of Hedge's *g* effects sizes as recommended by Larson-Hall (2016). Effect sizes show the strength of an effect, e.g., the mean difference. The point estimates of Cohen's *d* effect size are based on the pooled standard deviations as the standardiser. Effect sizes are based on the criteria set by Plonsky and Oswald (2014); the guidelines treat 0.4 as small, 0.7 as medium and 1 as large. Hedge's g is recommended by Gerlanc and Kirby (2013) and Larson-Hall (2016) as an unbiased and more conservative CI estimator. The results of the pilot studies and the aggregated data on the entire corpus will be discussed in detail in the following sections.

Bootstrapping was done using the *boot* package (version 1.3-20, Canty & Ripley, 2017). This code uses the BCa (Bias-Corrected and accelerated) bootstrapped confidence intervals as suggested by Larson-Hall (2016). The BCa method is useful as it corrects for the skewness and bias in the distribution of bootstrap estimates. The effect sizes of *d* and *g* were obtained from the *bootES* package (version 1.2; Gerlanc & Kirby, 2013).

To compensate for the possible increase of type I error rates as a result of multiple significance testing (e.g., one test with 22 lexical variables), the Bonferroni correction is applied to base the significance of any comparison on a stricter criterion. The new alpha level for the results of the final study and aggregated lexical tests was set to 0.05/22 = 0.002 and for the syntactic tests was set to 0.05/11= 0.004. The new alpha level for the results of the pilot study’s lexical dataset was set to 0.05/25 = 0.002, and for the syntactic dataset was set to 0.05/14 = 0.003.

The comparison of both genre-aggregated and genre-separated results for each measure (together with the rest of the tests in the following sections) clarify if the performance of students with regard to any measure is dependent on the rhetorical aspect of the text and the extent to which different rhetorical sections influence the lexical and syntactic values for each group of the students. In the following tables, the English L1 group is labelled as ‘NS’ that stands for Native Speaker of English. The asterisks in the tables in this chapter are printed only as visual aids. The codes for significance levels of all tables are as follows:

Annotation	p-value range	Significance level
***	[0, 0.001]	0.001
**	[0.001, 0.01]	0.01
*	[0.01, 0.05]	0.05

Tables 6.3 and 6.4 present the results of monofactorial tests and effect sizes of the statistically significant measures in the pilot study and aggregated lexical datasets respectively. Tables 6.5 and 6.6 present these results for the syntactic datasets. Both lexical and syntactic pilot studies included 150 abstracts, 50 per group. Both lexical and syntactic aggregated analyses were conducted on genre-aggregated datasets with 210 dissertations, 70 per group.

Table 6.3. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the pilot study lexical dataset

Pilot Study			**Tukey HSD group differences**			**ANOVA**		
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ls1	EFL 0.97 (0.02) ESL 0.96 (0.03) NS 0.95 (0.03)	ESL-EFL NS-EFL	-0.016 [-0.02, -0.00] -0.026 [-0.03, -0.01]	-0.79 -1.14	[-0.9, -0.2] [-1.32, -0.6]	11.35	<0.001	***
vs1	EFL 0.15 (0.04) ESL 0.18 (0.04) NS 0.18 (0.04)	ESL-EFL NS-EFL	0.02 [0.01, 0.04] 0.03 [0.01, 0.04]	0.68 0.71	[0.27, 1.09] [0.30, 1.09]	7.97	<0.001	***
vs2	EFL 0.97 (0.33) ESL 1.27 (0.47) NS 1.24 (0.34)	ESL-EFL NS-EFL	0.29 [0.14, 0.46] 0.26 [0.13, 0.39]	0.72 0.79	[0.34, 1.10] [0.35, 1.19]	8.9	<0.001	***
cvs1	EFL 0.68 (0.11) ESL 0.78 (0.14) NS 0.78 (0.10)	ESL-EFL NS-EFL	0.09 [0.04, 0.15] 0.09 [0.04, 0.13]	0.73 0.82	[0.31, 1.12] [0.37, 1.19]	9.64	<0.001	***
ttr	EFL 0.24 (0.03) ESL 0.26 (0.02) NS 0.26 (0.03)	ESL-EFL NS-EFL	0.02 [0.01, 0.03] 0.02 [0.01, 0.03]	0.77 0.71	[0.34, 1.14] [0.28, 1.10]	9.12	<0.001	***
logttr	EFL 0.78 (0.01) ESL 0.79 (0.01) NS 0.79 (0.01)	ESL-EFL	0.01 [0.00, 0.01]	0.8	[0.44, 1.14]	7.99	0.001	***

lv	EFL	0.10 (0.02)							
	ESL	0.11 (0.02)	NS-EFL	0.01 [0.00, 0.02]	0.7	[0.30, 1.09]	7.72	<0.001	*******
	NS	0.11 (0.02)							
vv1	EFL	0.16 (0.05)							
	ESL	0.20 (0.04)	ESL-EFL	0.03 [0.03, 0.05]	0.8	[0.28, 1.12]	8.82	<0.001	*******
	NS	0.20 (0.05)	NS-EFL	0.03 [0.01, 0.05]	0.9	[0.31, 1.16]			
svv1	EFL	1.08 (0.42)							
	ESL	1.47 (0.64)	ESL-EFL	0.39 [0.18, 0.60]	0.72	[0.3, 1.08]	9.24	<0.001	*******
	NS	1.47 (0.49)	NS-EFL	0.39 [0.21, 0.58]	0.87	[0.42, 1.23]			
cvv1	EFL	0.72 (0.13)							
	ESL	0.84 (0.18)	ESL-EFL	0.11 [0.05, 0.18]	0.76	[0.33, 1.08]	10.2	<0.001	*******
	NS	0.84 (0.14)	NS-EFL	0.12 [0.06, 0.17]	0.94	[0.46, 1.3]			
vv2	EFL	0.04 (0.01)							
	ESL	0.05 (0.01)	ESL-EFL	0.001 [0.001, 0.01]	0.64	[0.22, 0.99]	7.8	<0.001	*******
	NS	0.05 (0.01)	NS-EFL	0.01 [0.001, 0.01]	0.8	[0.36, 1.13]			

– Only the syntactic measures which showed between-group differences for at least one pair of comparison are included in this table. The non-significant results of measures and comparisons are provided in the link in supplementary materials/repository (see appendix B).

– The number of observations for all tests is 150. The degrees of freedom for all lexical analyses of variance are 2 and 147.

– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

Table 6.4. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the aggregated lexical dataset

Corpus				**Tukey HSD group differences**			**ANOVA**	
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ld	EFL 0.03 (0.01)	ESL-EFL	0.006 [0.003, 0.010]	0.70	[0.37, 0.99]	11.2	<.001	***
	ESL 0.04 (0.01)	NS-EFL	0.006 [0.002, 0.010]	0.77	[0.44, 1.11]			
	NS 0.04 (0.01)							
ls1	EFL 0.47 (0.10)							
	ESL 0.45 (0.15)	NS-EFL	-0.054 [-0.103,-0.005]	-0.50	[-0.85, -0.17]	3.49	0.03	*
	NS 0.42 (0.11)							
ndwesz	EFL 37.80 (1.09)							
	ESL 38.14 (1.39)	NS-EFL	0.53 [0.028, 1.032]	0.44	[0.10, 0.76]	3.21	0.04	*
	NS 38.33 (1.27)							
lv	EFL 0.39 (0.07)							
	ESL 0.36 (0.09)	ESL-EFL	-0.033 [-0.065,-0.001]	-0.40	[-0.74, -0.04]	3.63	0.02	*
	NS 0.36 (0.08)							
vv2	EFL 0.10 (0.03)	ESL-EFL	-0.018 [-0.029,-0.007]	-0.63	[-1, -0.29]			
	ESL 0.08 (0.03)	NS-EFL	-0.015 [-0.026,-0.004]	-0.62	[-0.94, -0.27]	9.35	<0.001	***
	NS 0.09 (0.02)							
nv	EFL 0.36 (0.07)	ESL-EFL	-0.039 [-0.070,-0.008]	-0.47	[-0.79, -0.13]			
	ESL 0.32 (0.09)	NS-EFL	-0.044 [-0.075,-0.013]	-0.65	[-0.95, -0.32]	6.89	0.001	**
	NS 0.32 (0.07)							

mattr	EFL 0.73 (0.02)	ESL-EFL	0.009[0.000, 0.018]	0.41	[0.03, 0.7]	8.30	<0.001	***
	ESL 0.74 (0.03)	NS-EFL	0.015 [0.006, 0.025]	0.75	[0.41, 1.07]			
	NS 0.74 (0.02)							
msttr	EFL 0.73 (0.02)							
	ESL 0.73 (0.03)	NS-EFL	0.016[0.006, 0.025]	0.77	[0.44, 1.11]	8.48	<0.001	***
	NS 0.74 (0.02)							
mtld	EFL 50.94 (4.47)							
	ESL 53.98 (9.86)	NS-EFL	5.23 [2.02, 8.44]	0.75	[0.42, 1.05]	7.47	<0.001	***
	NS 56.19 (8.77)							

– Only the syntactic measures which showed between-group differences for at least one pair of comparison are included in this table. The non-significant results of measures and comparisons are provided in the link in supplementary materials/repository (see appendix B).
– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

Nine out of the twenty-two lexical measures showed significant between-group differences in the aggregated dataset; amongst them, the ld, vv2, mattr, msttr, and mtld indices captured more between-group differences satisfying the stricter criterion of the Bonferroni-corrected new alpha level of 0.002.

Similar to the results of the pilot study (with abstracts only), this final analysis on the whole corpus also point out to similar performances of the ESL and English L1 groups and significant differences are found only in pair-wise comparisons which include the EFL group. Since the pilot study was conducted to aid the measure-selection process, I will not include a detailed analysis of the results here and focus mainly on the results of the main study. However, a few noteworthy points are discussed below.

In contrast to the findings of the pilot study, the EFL group outperformed the ESL group in the values of lv, vv2, and nv measures which belong to the category of lexical variation of word classes, and outperformed the English L1 group in the values of ls1, vv2, and nv measures. Since the pilot study was conducted on 50 abstracts from each group only, we can clearly see the impact of the increase in the sample size as well as the total impact of all rhetorical sections of the dissertations (the entire dissertations as opposed to one section) on the number and type of lexical measures which show between-group differences. With a smaller sample size in the pilot study, eleven lexical measures (out of 25) captured the differences between at least one pair of group comparisons with medium to large effect sizes; the final study with a larger sample size on the aggregated corpus resulted in five lexical measures showing the between-group differences which passed the stricter significance level.

The results of the rest of the measures which marked between-group differences on the aggregated lexical data (i.e., ld, ndwesz, mattr, msttr, and mtld) are a testament to the outperformance of the English L1 and ESL groups in producing lexically diverse texts. This group of measures all calculate the word strings or segments suggesting that this group of lexical diversity indices are different from the group of ratio-based word-class lexical diversity (this claim is further supported in section 6.3.1 and the results of factor loadings in the table 6.27). The fact that this word-string-based group of lexical diversity produced narrower Cis (closer to the mean estimate) than the group of TTR of word classes, also indicate that they are better indicators of performance differences with regard to lexically diverse texts; i.e., they can reduce the effect of the increase in the number of tokens more effectively.

Table 6.5. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the pilot study syntactic dataset

Pilot Study		**Tukey HSD group differences**				**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLS	EFL 24.10 (5.21)							
	ESL 25.91 (6.09)	NS-EFL	3.26 [0.3, 6.22]	0.51	[0.13, 0.87]	3.43	0.03	*
	NS 27.23 (7.27)							
MLT	EFL 21.62 (4.95)							
	ESL 23.99 (5.55)	NS-EFL	3.8 [1.05, 6.56]	0.64	[0.25, 0.98]	5.47	0.005	**
	NS 25.43 (6.79)							
C/T	EFL 1.59 (0.26)							
	ESL 1.72 (0.32)	NS-EFL	0.23 [0.07, 0.38]	0.71	[0.32, 1.08]	6.5	0.001	***
	NS 1.82 (0.39)							
CT/T	EFL 0.38 (0.16)							
	ESL 0.47 (0.18)	NS-EFL	0.15 [0.06, 0.24]	0.79	[0.4, 1.18]	8.44	<.001	***
	NS 0.54 (0.23)							
DC/C	EFL 0.29 (0.11)							
	ESL 0.35 (0.11)	ESL-EFL	0.05 [0.001, 0.11]	0.53	[0.13, 0.93]	7.75	<.001	***
	NS 0.38 (0.13)	NS-EFL	0.09 [0.03, 0.14]	0.74	[0.33, 1.16]			
DC/T	EFL 0.48 (0.23)							
	ESL 0.63 (0.28)	NS-EFL	0.25 [0.11, 0.4]	0.8	[0.42, 1.14]	8.79	<.001	***
	NS 0.74 (0.4)							

CN/T	EFL 3.16 (0.87) ESL 3.48 (0.98) NS 3.86 (1.18)	NS-EFL	0.69 [0.21, 1.17]	0.66	[0.26, 1.04]	5.83	0.003	***
VP/T	EFL 2.19 (0.45) ESL 2.44 (0.56) NS 2.57 (0.76)	NS-EFL	0.37 [0.09, 0.66]	0.6	[0.22, 0.94]	5.04	0.007	**

– Only the syntactic measures which showed between-group differences for at least one pair of comparison are included in this table. The non-significant results of measures and comparisons are provided in the link in supplementary materials/repository (see appendix B).
– The number of observations for all tests is 150. The degrees of freedom for all lexical analyses of variance are 2 and 147.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.003.

Table 6.6. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the aggregated syntactic dataset

Corpus		**Tukey HSD group differences**				**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 21.18 (2.12) ESL 23.87 (3.93) NS 22.42 (3.19)	ESL-EFL NS-EFL	2.68 [1.42, 3.95] -1.45 [-2.72, -0.19]	0.85 0.45	[0.53, 1.15] [0.12, 0.77]	12.64	<.001	***
C/T	EFL 1.73 (0.12) ESL 1.91 (0.22) NS 1.87 (0.23)	ESL-EFL NS-EFL	0.17 [0.09, 0.25] 0.13 [0.05, 0.21]	0.97 0.71	[0.65, 1.28] [0.38, 1]	14.94	<.001	***
CT/T	EFL 0.44 (0.05) ESL 0.52 (0.08) NS 0.51 (0.09)	ESL-EFL NS-EFL	0.079 [0.049, 0.11] 0.076 [0.046, 0.107]	1.15 1.04	[0.77, 1.49] [0.69, 1.39]	24.33	<.001	***
DC/C	EFL 0.36 (0.04) ESL 0.41 (0.06) NS 0.40 (0.06)	ESL-EFL NS-EFL	0.049 [0.029, 0.07] 0.043 [0.02, 0.06]	1.03 0.9	[0.68, 1.35] [0.53, 1.2]	19.29	<.001	***
DC/T	EFL 0.63 (0.10) ESL 0.80 (0.20) NS 0.77 (0.20)	ESL-EFL NS-EFL	0.16 [0.09, 0.23] 0.13 [0.06, 0.20]	1.02 0.83	[0.7, 1.32] [0.49, 1.12]	17.41	<.001	***
CP/C	EFL 0.42 (0.08) ESL 0.38 (0.10) NS 0.36 (0.09)	ESL-EFL NS-EFL	-0.04 [-0.07, -0.004] -0.05 [-0.08, -0.01]	-0.45 -0.61	[-0.79, -0.09] [-0.96, -0.21]	6.38	0.002	**

CN/T	EFL 2.99 (0.35)	ESL-EFL	0.32 [0.107, 0.54]	0.6	[0.26, 0.89]	6.64	0.001	**
	ESL 3.32 (0.68)	NS-ESL	-0.23 [-0.45, -0.019]	0.19	[-0.13, 0.5]			
	NS 3.08 (0.56)							
VP/T	EFL 2.32 (0.22)	ESL-EFL	0.24 [0.11, 0.37]	0.79	[0.43, 1.07]	10.14	<.001	***
	ESL 2.57 (0.38)	NS-EFL	0.16 [0.029, 0.29]	0.53	[0.21, 0.84]			
	NS 2.48 (0.36)							

– Only the syntactic measures which showed between-group differences for at least one pair of comparison are included in this table. The non-significant results of measures and comparisons are provided in the link in supplementary materials/repository (see appendix B).
– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

The findings for the aggregated syntactic data show statistically significant values for mean differences, also as indicated by the CIs and effect sizes of eight syntactic measures, each for two sets of group comparisons. The largest effect size (d = 1.15) is marked for the CT/T index that measures the ratio of complex T-units to all T-units. The analysis of the aggregated syntactic data also shows that the measures with T-unit in the denominator received larger effect sizes, suggesting that group differences in the aggregated data are more illustrative using T-unit.

Eight of the 11 syntactic complexity measures show significant between-group differences in at least two comparison pairs at or below the new Bonferroni-corrected alpha level of 0.004, as demonstrated in table 6.6. Seven of these measures consistently show differences between the ESL-EFL and NS-EFL comparison sets, highlighting the ESL and English L1 groups' more syntactically complex dissertations in all four constructs of syntactic complexity as outlined in the previous chapter. This similar performance of the English L1 and ESL groups will be examined and discussed again regarding the individual rhetorical sections.

Similar to the results of the lexical analysis on the aggregated dataset, the syntactic complexity of the entire dissertations is shown to be higher for the ESL and English L1 groups, with the ESL group showing larger values than the English L1 group in the length of T-units and the number of complex nominals.

The EFL group's dissertations are also shown to be less syntactically complex than the other two groups, except the number of coordinate phrases as calculated via the CP.C (coordinate phrases per clause) index. This, as will be discussed in detail in the syntactic analyses of individual rhetorical sections and with reference to the previous research, could be an indicator of lower syntactic proficiency.

Regarding the results of between-group differences in the syntactic pilot study, five measures showed statistically significant differences (based on the alpha level of 0.003) for the NS-EFL comparison set suggesting that the EFL group produced the least-syntactically-complex abstracts. The measures that captured these differences were MLS (mean length of sentence), MLT (mean length of T-units), VP/T (verb phrases per T-unit), C/T (T-unit complexity ratio), DC/C and DC/T (dependent clauses per clause and per T-unit), CT/T (complex T-units per T-unit or complex T-unit ratio), and CN/T (complex nominals per T-unit). This indicates that the EFL group produced a relatively smaller amount of subordination , especially dependent clauses as well as shorter sentences and T-units, and an overall lower phrasal complexity than the other two groups. The results for MLT and MLS were almost identical; this was one of the reasons I dropped the MLS index from the final set of measures for the final analyses. Once more, the ESL and English L1 groups performed very similarly regarding the production of syntactically complex texts as measured via the 14 indices available in L2SCA.

6.3.1. Lexical Complexity in Six Rhetorical Sections of Dissertations

Tables 6.7 to 6.12 demonstrate the findings of the mentioned statistics for the lexical measures which showed significant results (between-group differences) in each of the six rhetorical sections of students' dissertations. The significant differences are based on the new Bonferroni-corrected alpha level of 0.002 for the lexical dataset. Only the measures which showed significant differences in comparisons will be printed in the tables. The link to the non-significant results of measures and comparisons will be provided in Appendix B. In the absence of previous research on the analysis of dissertations' rhetorical sections regarding these sets of complexity measures, I discuss the results based on other academic and SLA studies that examined written or spoken corpora using any of the complexity measures investigated in this study.

Table 6.7 presents the results of significant between-group differences in the abstracts section of the final study. The comparison of the findings of the pilot study with 50 abstracts in each group (table 6.3) and the final study with 70 **abstracts** in each group also reveals interesting differences in the number and the type of lexical measures which showed between-group differences. The final study with a larger sample size led to a greater number of lexical measures indicating significant group differences with larger effect sizes; besides, no lexical sophistication index was spotted among them (table 6.7). Regarding the type of pair-wise comparisons, the ESL and English L1 groups again performed similarly in the production of lexically diverse texts, and once again, the differences only involve the EFL group. Apart from the maas index (a logarithm-based measure), the rest of the measures as indicated in table 6.7 mark larger values for the English L1 and ESL groups, with some measures like ndwesz (number of different words, type I), mattr (moving-average type-token ratio), msttr (mean segmental type-token ratio), hdd (hypergeometric D), mtld (measure of textual lexical diversity), and vocd (the original D measure) recording medium to large effect sizes for the ESL-EFL and NS-EFL mean differences.

Table 6.7. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the abstract sections on the lexical dataset

Abstract		**Tukey HSD group differences**				**ANOVA**		
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ld	EFL 0.02 (0.02)	ESL-EFL	0.011 [0.002,0.019]	0.52	[0.20, 0.80]	5.72	0.003	**
	ESL 0.04 (0.02)	NS-EFL	0.009 [0.001, 0.017]	0.51	[0.16, 0.81]			
	NS 0.03 (0.02)							
ndwerz	EFL 37.35 (1.93)	ESL-EFL	1.257 [0.499, 2.014]	0.67	[0.30, 1.03]	9.62	0.000	***
	ESL 38.61 (1.77)	NS-EFL	1.177 [0.419, 1.934]	0.60	[0.25, 0.96]			
	NS 38.53 (1.98)							
ndwesz	EFL 36.83 (2.20)	ESL-EFL	1.742 [0.857, 2.628]	0.81	[0.39, 1.15]	12.12	<.001	***
	ESL 38.58 (2.08)	NS-EFL	1.401 [0.515, 2.287]	0.61	[0.23, 0.96]			
	NS 38.24 (2.37)							
rttr	EFL 7.45 (0.78)	ESL-EFL	0.364 [0.049, 0.679]	0.47	[0.09, 0.79]	5.14	0.006	**
	ESL 7.82 (0.75)	NS-EFL	0.377 [0.062, 0.692]	0.46	[0.09, 0.79]			
	NS 7.83 (0.84)							
logttr	EFL 0.87 (0.02)	ESL-EFL	0.010 [0.002, 0.018]	0.54	[0.14, 0.89]	6.60	0.001	**
	ESL 0.88 (0.02)	NS-EFL	0.010 [0.002, 0.018]	0.54	[0.19, 0.86]			
	NS 0.88 (0.02)							
uber	EFL 18.53 (2.40)	ESL-EFL	1.662 [0.538, 2.875]	0.65	[0.28, 1.01]	9.39	<.001	***
	ESL 20.19 (2.66)	NS-EFL	1.887 [0.764, 3.010]	0.65	[0.33, 0.96]			
	NS 20.42 (3.31)							
mass	EFL 0.06 (0.01)	ESL-EFL	-0.006[-0.009, -0.003]	-0.78	[-1.10, -0.30]	15.83	<.001	***

	ESL 0.05 (0.01)	NS-EFL	-0.007 [0.010, -0.003]	-0.84	[-1.15, -0.50]			
	NS 0.05 (0.01)							
mattr	EFL 0.71 (0.04)	ESL-EFL	0.035 [0.018, 0.051]	0.87	[0.49, 1.18]	18.76	<.001	*******
	ESL 0.74 (0.04)	NS-EFL	0.038, 0.022, 0.055]	0.92	[0.52, 1.30]			
	NS 0.75 (0.04)							
msttr	EFL 0.71 (0.04)	ESL-EFL	0.031 [0.014, 0.047]	0.72	[0.37, 1.06]	20.37	<.001	*******
	ESL 0.74 (0.04)	NS-EFL	0.043 [0.026, 0.059]	1.04	[0.68, 1.41]			
	NS 0.75 (0.04)							
hdd	EFL 0.75 (0.04)	ESL-EFL	0.029 [0.015, 0.043]	0.81	[0.41, 1.16]	16.94	<.001	*******
	ESL 0.78 (0.04)	NS-EFL	0.030 [0.016, 0.044]	0.87	[0.48, 1.22]			
	NS 0.78 (0.03)							
mtld	EFL 48.36 (10.84)	ESL-EFL	10.55 [5.31, 15.78]	0.86	[0.52, 1.18]	17.49	<.001	*******
	ESL 58.92 (13.34)	NS-EFL	12.02 [6.78, 17. 26]	0.92	[0.55, 1.21]			
	NS 60.39 (14.87)							
vocd	EFL 65.89 (16.07)	ESL-EFL	12.42 [5.17, 19.67]	0.70	[0.34, 1.02]	12.41	<.001	*******
	ESL 78.32 (19.05)	NS-EFL	13.92 [6.68, 21.17]	0.78	[0.42, 1.11]			
	NS 79.82 (19.18)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

The ld (lexical density) index which did not point out to any between-group differences in the pilot study, did show differences for two sets of group comparisons in the final analysis, showing that English L1 and ESL groups produced more lexically-dense abstracts. This finding suggests that with large enough sample, the possibility of finding group differences in the values of lexical density increases.

The same holds true for the uber (a logarithm-based measure) and rttr (root TTR) indices which did not reveal any pair-wise differences in the pilot study but did show medium-size effects in the differences between the NS-EFL and ESL-EFL sets in the final analysis with a larger sample. As for the comparison of the aggregated corpus vs. the abstract section of the final study, the lexical variations based on the TTR of word classes are absent from the abstract section's significant results, while both datasets share the ndwesz, mattr, msttr, and mtld measures as the ones with larger effects sizes in spotting the group differences.

The findings of the **introduction** rhetorical section (table 6.8) point to some similarities with the abstract section of the final study: in both datasets, ndwesz, ndwerz, logttr, uber, mattr, msttr, mtld and vocd consistently show differences for the NS-EFL and ESL-EFL comparison sets with medium to large effect sizes and the p-values of the F statistic satisfying the stricter Bonferroni-corrected criterion of 0.002. Similar to previous findings, the EFL group produced less-lexically-diverse texts than the English L1 and ESL groups.

In both sections, it is the construct of lexical diversity that is the dominant construct for distinguishing the abstracts and introduction sections of M.A dissertations of the three groups. Lexical density in the abstract sections and lexical sophistication in the introduction sections only show small effects for the differences.

The measures capturing significant between-group differences that are shared between the aggregated corpus and the introduction section are mattr, msttr, and mtld. The two indices of ndwerz and ndwesz which calculate the number of different words, show significant differences with medium to large effect sizes between the mentioned two sets of group comparisons in both the abstract and the introduction sections of the non-aggregated data, but did not show any differences in the aggregated corpus, nor in any other rhetorical sections. This suggests that these indices' values are highly dependent on text length as well as on the rhetorical section or the sub-genre of the academic writing: the two relatively-shorter rhetorical sections of abstract and introduction where the disproportionate effect of the number of tokens on the sub-samples are reduced, contain a greater number of different words.

Table 6.8. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the introduction sections on the lexical dataset

Introduction			**Tukey HSD group differences**			**ANOVA**		
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ls2	EFL 0.34 (0.06)	ESL-EFL	0.024 [0.001, 0.048]	0.42	[0.07, 0.73]	4.30	0.014	*
	ESL 0.37 (0.06)	NS-EFL	0.025 [0.002, 0.048]	0.43	[0.09, 0.76]			
	NS 0.37 (0.06)							
ndwerz	EFL 39.23 (1.56)	ESL-EFL	0.848 [0.283, 1.414]	0.57	[0.23, 0.91]	7.15	<.001	*******
	ESL 40.08 (1.37)	NS-EFL	0.700 [0.134, 1.265]	0.48	[0.13, 0.79]			
	NS 39.93 (1.31)							
ndwesz	EFL 37.62 (2.06)	ESL-EFL	1.280 [0.575, 1.984]	0.69	[0.34, 0.98]	15.06	<.001	*******
	ESL 38.90 (1.61)	NS-EFL	1.525 [0.820, 2.230]	0.82	[0.51, 1.14]			
	NS 39.14 (1.59)							
logttr	EFL 0.84 (0.01)	ESL-EFL	0.008 [0.0009, 0.0161]	0.45	[0.13, 0.76]	4.61	0.010	*
	ESL 0.85 (0.02)	NS-EFL	0.008 [0.0008, 0.0160]	0.49	[0.17, 0.81]			
	NS 0.85 (0.02)							
uber	EFL 19.64 (1.59)	ESL-EFL	1.001 [0.218, 1.784]	0.53	[0.20, 0.84]	5.90	0.003	**
	ESL 20.64 (2.12)	NS-EFL	0.971 [0.188, 1.754]	0.51	[0.22, 0.80]			
	NS 20.61 (2.12)							
adjv	EFL 0.04 (0.03)							*
	ESL 0.05 (0.04)	NS-EFL	0.018 [0.000, 0.035]	0.42	[0.10, 0.74]	3.85	0.022	
	NS 0.06 (0.05)							

maas	EFL 0.05 (0.00)	ESL-EFL	-0.002 [-0.004, -2.528]	-0.44	[-0.75, -0.09]	4.25	0.015	*
	ESL 0.05 (0.01)							
	NS 0.05 (0.00)							
mattr	EFL 0.73 (0.03)	ESL-EFL	0.016 [0.005, 0.027]	0.56	[0.23, 0.86]	12.43	<.001	***
	ESL 0.75 (0.03)	NS-EFL	0.022 [0.011, 0.033]	0.87	[0.55, 1.15]			
	NS 0.75 90.03)							
msttr	EFL 0.73 (0.03)	ESL-EFL	0.014 [0.003, 0.026]	0.50	[0.18, 0.81]	10.21	<.001	***
	ESL 0.75 (0.03)	NS-EFL	0.021 [0.009, 0.032]	0.80	[0.42, 1.10]			
	NS 0.75 (0.03)							
mtld	EFL 52.52 (8.27)	ESL-EFL	5.494 [1.190, 9.797]	0.49	[0.16, 0.77]	7.34	<.001	***
	ESL 58.01 (13.53)	NS-EFL	6.485 [2.182, 10.78]	0.71	[0.38, 1.02]			
	NS 59.01 (9.87)							
vocd	EFL 85.46 (14.91)	ESL-EFL	9.49 [2.728, 16.26]	0.57	[0.20, 0.88]	8.45	<.001	***
	ESL 94.96 (18.01)	NS-EFL	10.80 [4.03, 17.57]	0.68	[0.33, 0.98]			
	NS 96.26 (17.79)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.

– The significant results of ANOVA shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

Table 6.9. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the literature review sections on the lexical dataset

Lit. Review		Tukey HSD group differences				ANOVA		
Lexical Measures	Mean and (SD)	Group Comparisons	Mean difference & [95% BCa CIs]	Effect size Cohen's d	Bootstrapped Effect size [95% BCa CIs] Hedges' g	F	Pr(>F)	Sig.
ld	EFL 0.03 (0.01)	ESL-EFL	0.0058 [0.001, 0.010]	0.52	[0.21, 0.83]	6.20	0.002	**
	ESL 0.04 (0.01)	NS-EFL	0.0057 [0.001, 0.010]	0.57	[0.24, 0.87]			
	NS 0.04 (0.01)							
rttr	EFL 15.16 (1.48)	ESL-EFL	-0.886 [-1.642, -0.130]	-0.50	[-0.88, -0.14]	4.52	0.011	*
	ESL 14.27 (2.02)	NS-EFL	-0.770 [-1.526, -0.014]	-0.42	[-0.73, -0.03]			
	NS 14.39 (2.12)							
logttr	EFL 0.82 (0.01)	ESL-EFL	-0.008 [-0.015, -1.950]	-0.50	[-0.84, -0.15]	5.04	0.007	**
	ESL 0.81 (0.02)							
	NS 0.82 (0.02)							
uber	EFL 20.72 (1.04)	ESL-EFL	-0.840 [-1.427,- 0.253]	-0.60	[-0.96, -0.24]	6.99	0.001	**
	ESL 19.88 (1.66)	NS-EFL	-0.764 [-1.351, -0.177]	-0.55	[-0.89, -0.17]			
	NS 19.96 (1.63)							
lv	EFL 0.39 (0.10)	ESL-EFL	-0.061 [-0.101, -0.021]	-0.58	[-0.93, -0.19]	8.62	<.001	***
	ESL 0.33 (0.11)	NS-EFL	-0.059 [-0.099, -0.019]	-0.62	[-0.94, -0.28]			
	NS 0.33 (0.09)							
vv2	EFL 0.11 (0.04)	ESL-EFL	-0.028 [-0.042, -0.015]	-0.80	[-1.12, -0.45]	15.34	<.001	***
	ESL 0.08 (0.03)	NS-EFL	-0.025 [-0.038, -0.012]	-0.73	[-1.06, -0.38]			

	NS 0.08 (0.03)							
nv	EFL 0.36 (0.10)	ESL-EFL	-0.064 [-0.105, -0.022]	-0.56	[-0.93, -0.15]	9.16	<.001	***
	ESL 0.29 (0.13)	NS-EFL	-0.066 [-0.107, -0.024]	-0.72	[-1.05, -0.38]			
	NS 0.29 (0.08)							
maas	EFL 0.05 (0.00)	ESL-EFL	0.002 [8.187, 0.003]	0.64	[0.28, 1.04]	6.19	0.001	**
	ESL 0.05 (0.00)	NS-EFL	0.001 [4.729, 0.003]	0.45	[0.09, 0.79]			
	NS 0.05 (0.00)							
hdd	EFL 0.82 (0.01)	ESL-EFL	-0.008 [-0.015, -0.001]	-0.47	[-0.81, -0.14]	4.41	0.013	*
	ESL 0.81 (0.02)							
	NS 0.81 (0.02)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

Other indices like ls2 (lexical sophistication type II), adjv (adjective variation) and maas (a logarithm-based measure) are only significant at the 0.05 level and not at the stricter 0.002 level.

Considering the group comparisons in the **literature review** sections of the dissertations (table 6.9), we see that the EFL group is surprisingly producing more lexically-diverse texts than the other two groups as indicated by the values of most of the indices which show group differences (i.e., the seven indices of rttr (root TTR), logttr (logarithmic TTR), uber (another logarithm-based measure of lexical diversity), lv (lexical variation), vv2 (verb variation type II), nv (noun variation), and hdd (a variant of D measure)).

This finding is noteworthy since they are the other two groups that are performing better in the syntactic indices' values in the same section of literature review (table 6.16); so the higher lexical values of the EFL group are due to the increased rate of producing new and varied vocabulary. However, it should be noted that these group differences in the lexical indices have small to medium effect sizes which indicate a relative outperformance of the EFL group and not a substantial one. A quick look at these indices also shows that apart from the hdd index (with a small effect size for the ESL-EFL comparison), the rest of these measures are based on TTR ratios.

Type-token ratio-based measures are sensitive to the text length and the increase in the number of tokens. Even the logttr measure which reduces the effect of the increasing number of tokens, is only showing a significant difference at the level of 0.007 which does not satisfy the stricter criterion of 0.002 as set by Bonferroni correction. Rttr which also reduces the effect of the increasing number of tokens by taking their square root, produces significant between-group differences only at the 0.01 level. Among this group of indices, only the four measures of uber, lv, nv, and vv2 are significant at the strict 0.002 level, proving a genuine outperformance of the EFL group in the production of the new and diverse lexical verb and noun types at a higher rate.

Regarding important lexical constructs, once more lexical diversity is shown to be the dominant construct for distinguishing the literature review rhetorical sections of the three groups. As mentioned, this distinction is more noticeable in the use of varied nouns and verbs for the EFL group. This interesting finding will be revisited in the discussion of the key points of this chapter.

The analysis of the **method and design** rhetorical sections of the dissertations (table 6.10) also reveals similar results for the nv and vv2 indices with small effects on the group comparisons, showing that the EFL group produced more verb and noun types.

Table 6.10. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the method & design sections on the lexical dataset

Method		**Tukey HSD group differences**				**ANOVA**		
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ld	EFL 0.03 (0.01)	ESL-EFL	0.007 [0.002, 0.011]	0.64	[0.30, 0.96]	11.17	<.001	***
	ESL 0.03 (0.01)	NS-EFL	0.008 [0.003, 0.012]	0.79	[0.39, 1.12]			
	NS 0.03 (0.01)							
ls1	EFL 0.52 (0.12)	ESL-EFL	-0.057 [-0.111, -0.002]	-0.41	[-0.76, -0.07]	9.42	<.001	***
	ESL 0.46 (0.15)	NS-EFL	-0.099 [-0.154, -0.045]	-0.78	[-1.14, -0.37]			
	NS 0.42 (0.13)							
ndwesz	EFL 37.04 (1.37)					4.18	0.016	*
	ESL 37.70 (2.08)	NS-EFL	0.825 [0.113, 1.538]	0.50	[0.14, 0.84]			
	NS 37.87 (1.84)							
vv2	EFL 0.14 (0.07)	ESL-EFL	-0.029 [-0.056, -0.002]	-0.41	[-0.78, -0.05]	3.78	0.024	*
	ESL 0.11 (0.08)							
	NS 0.12 90.06)							
nv	EFL 0.47 (0.16)	ESL-EFL	-0.069 [-0.134, -0.003]	-0.40	[-0.73, -0.03]	4.78	0.009	**
	ESL 0.40 (0.18)	NS-EFL	-0.078 [-0.144, -0.013]	-0.50	[-0.84, -0.15]			
	NS 0.39 (0.15)							
mattr	EFL 0.71 (0.02)	ESL-EFL	0.012 [0.001, 0.023]	0.44	[0.11, 0.75]	8.22	<.001	***
	ESL 0.72 (0.03)	NS-EFL	0.018 [0.007, 0.030]	0.76	[0.39, 1.14]			
	NS 0.73 (0.03)							

msttr	EFL 0.71 (0.02)	ESL-EFL	0.012 [0.001, 0.023]	0.42	[0.11, 0.75]	8.03	<.001	***
	ESL 0.72 (0.03)	NS-EFL	0.018 [0.007, 0.030]	0.73	[0.34, 1.07]			
	NS 0.73 (0.03)							
hdd	EFL 0.78 (0.02)					4.79	0.009	**
	ESL 0.79 (0.03)	NS-EFL	0.011 [0.002, 0.019]	0.57	[0.21, 0.89]			
	NS 0.79 (0.02)							
mtld	EFL 45.81 (6.71)	ESL-EFL	4.343 [0.522, 8.164]	0.45	[0.12, 0.71]	7.60	<.001	***
	ESL 50.15 (11.87)	NS-EFL	6.139 [2.318, 9.960]	0.75	[0.39, 1.03]			
	NS 51.95 (9.45)							
vocd	EFL 87.38 (14.32)	ESL-EFL	7.961 [1.262, 14.66]	0.44	[0.10, 0.74]	5.48	0.004	**
	ESL 95.34 (20.74)	NS-EFL	8.312 [1.613, 15.01]	0.57	[0.19, 0.91]			
	NS 95.69 (14.50)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the new Bonferroni-corrected alpha level of 0.002.

Table 6.11. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the results & discussion sections on the lexical dataset

Results				**Tukey HSD group differences**			**ANOVA**	
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ld	EFL 0.03 (0.01)	ESL-EFL	0.009 [0.003, 0.014]	0.68	[0.32, 0.94]	9.88	<.001	***
	ESL 0.04 (0.02)	NS-EFL	0.008 [0.003, 0.013]	0.69	[0.37, 1.03]			
	NS 0.04 (0.01)							
ls1	EFL 0.50 (0.14)					6.46	0.001	**
	ESL 0.44 (0.19)	NS-EFL	-0.096 [-0.160, -0.032]	-0.66	[-1.02, -0.28]			
	NS 0.41 (0.15)							
ndwesz	EFL 36.83 (2.24)					4.55	0.011	*
	ESL 37.52 (1.95)	NS-EFL	0.955 [0.183, 1.727]	0.49	[0.16, 0.79]			
	NS 37.78 (1.56)							
vv1	EFL 0.32 (0.16)	ESL-EFL	0.075 [0.003, 0.147]	0.42	[0.06, 0.75]	3.22	0.041	*
	ESL 0.40 (0.19)							
	NS 0.38 (0.19)							
vv2	EFL 0.08 (0.03)	ESL-EFL	-0.014 [-0.028, -0.001]	-0.43	[-0.79, -0.07]	3.96	0.020	*
	ESL 0.07 (0.04)							
	NS 0.07 (0.03)							
mattr	EFL 0.70 (0.04)	ESL-EFL	0.021 [0.006, 0.035]	0.56	[0.26, 0.84]	10.29	<.001	***
	ESL 0.72 (0.03)	NS-EFL	0.025 [0.011, 0.040]	0.69	[0.37, 0.95]			

	NS 0.73 (0.03)							
msttr	EFL 0.70 (0.04)	ESL-EFL	0.020 [0.006, 0.034]	0.54	[0.22, 0.81]	9.51	<.001	***
	ESL 0.72 (0.03)	NS-EFL	0.024 [0.010, 0.039]	0.67	[0.36, 0.94]			
	NS 0.73 90.03)							
hdd	EFL 0.78 (0.03)	ESL-EFL	0.012 [0.000, 0.024]	0.38	[0.06, 0.67]	3.92	0.021	*
	ESL 0.79 (0.03)	NS-EFL	0.011 [0.000, 0.023]	0.40	[0.10, 0.69]			
	NS 0.79 (0.02)							
mtld	EFL 45.19 (7.76)	ESL-EFL	5.233 [1.396, 9.070]	0.56	[0.24, 0.89]	9.08	<.001	***
	ESL 50.42 (10.71)	NS-EFL	6.547 [2.710, 10.38]	0.72	[0.38, 1.02]			
	NS 51.74 (10.12)							
vocd	EFL 87.63 (15.87)	ESL-EFL	8.850 [1.151, 16.549]	0.42	[0.12, 0.71]	3.68	0.026	*
	ESL 96.48 (24.50)							
	NS 92.33 (16.28)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the Bonferroni-corrected alpha level of 0.002.

Sophisticated lexical tokens to the number of lexical tokens as named with ls1, is another instance where the EFL students outperformed the other two groups with medium effect sizes, but the p-value shows the probability of the F statistic (F = 9.42) due to chance is <.001.

Since the sophisticated lexical items in this study are filtered through both BNC and BAWE frequently-used word lists, larger mean difference and CI values for the EFL group suggest that this group employed more sophisticated and infrequent words in the method section, as well as in the results and conclusion rhetorical sections as presented in tables 6.11 and 6.12. Concerning the rest of measures (i.e., ld, mattr, msttr, and mtld), the ESL and English L1 groups produced more lexically-complex texts with significant differences with the EFL group. These differences mark medium effect sizes. The p-values of the ANOVA tests also pass the stringent Bonferroni-corrected level. This pattern is so far consistent in the analyses of the rhetorical sections: the lexical diversity measures based on the TTR of word classes showing larger mean values for the EFL group and the word-string-based lexical diversity measures showing larger mean values for the English L1 and ESL groups. Considering the fact that the former types of indices use content/lexical words (e.g., nouns, verbs) and the word-string-based indices calculate all words, we notice the effect of function words as well as the effect of the quantification methods (e.g., ratio-based vs. word segments) on these group differences. Hdd and vocd indices being similar in the computation process produced similar results also regarding the type of group comparisons and the sizes of their effects on such comparisons (table 6.10).

Lexical density values show very similar patterns in the aggregated corpus and the method section both in terms of the group comparisons as well as the mean difference, confidence intervals, and the significance tests' values. This pattern is repeated for the next rhetorical sections of results and conclusion, as presented in tables 6.11 and 6.12 and as will be discussed in the following paragraphs. In section 6.6.1, I will revisit these findings based on the results of the interaction effects of groups and rhetorical sections. The combined results suggest that the descriptive and reporting rhetorical sections like abstract, result, and conclusion are more lexically dense especially in the English L1 and ESL students' texts than the explanatory and informational rhetorical sections like introduction and literature review. The text-length dependency of lexical density, however, cannot be supported as the token counts of the results section is similar to the literature review for all groups.

The method & design sections witness the presence of all three constructs of lexical complexity with similar effects for distinguishing these groups' texts. Overall, the texts of

English L1 and ESL groups seem to be more lexically dense and diverse, whereas the EFL group's texts have larger numbers of sophisticated words as filtered against a general and a field-specific word list.

The mean difference values of the ls1 (lexical sophistication type I) and vv2 (verb variation type II) measures in the **results and discussion** rhetorical sections (table 6.11) resemble those of the method and design rhetorical section: these values once more point out the outperformance of the EFL group, albeit with a small effect for the vv2 and medium effect size for lexical sophistication type 1. Similar studies need to be conducted to rule out the possible effects of sub-disciplinary variations regarding the use of sophisticated terms and to examine whether EFL academic writers, e.g., in Iran or elsewhere genuinely outperform the English L1s regarding the amount of sophisticated lexical items based on external word lists as reference points. This point will be partially examined in the linguistic examples from the texts of the three groups in 6.3.2.

Mattr, msttr, and mtld measures again captured significant between-group differences at the 0.002 level with medium Cohens' *d* effect sizes and hedge's *g* effect size confidence intervals which reach up to 0.9 and 1. These three indices all belong to the lexical diversity of word strings/segments. The significant group comparisons for these three indices, as with the findings of previous rhetorical sections, denote the lexical complexity of the texts of the English L1, ESL, and EFL groups respectively. Likewise, lexical density mean-difference values suggest the outperformance of the English L1 and ESL groups, with medium effects and a large F statistic. The comparison of the results section with the aggregated data also reveals similar findings for the lexical density, ls1 and ndwesz both in terms of the type and number of group comparisons, and the significance levels. Other measures which showed between-group differences in both datasets are vv2 (two significant comparisons for the aggregated data and one for the results section), mattr, msttr and mtld (with two significant comparisons for the results section and one for the aggregated data).

Regarding the overall important constructs that can distinguish the texts of the three groups, the results & discussion rhetorical sections show a similar profile to the method & design sections with the similar presence of all three constructs and similar effects. However, unlike the previous rhetorical section, in this section, we notice mixed results regarding the two verb-based indices of lexical variation for the ESL-EFL comparison set in that vv1 shows the ESL text's more use of varied verbs but the vv2 index showing the opposite.

Table 6.12. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the conclusion sections on the lexical dataset

Conclusion		**Tukey HSD group differences**				**ANOVA**		
Lexical Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
ld	EFL 0.03 (0.01)	ESL-EFL	0.006 [0.000, 0.011]	0.46	[0.11, 0.78]	8.61	<.001	***
	ESL 0.03 (0.01)	NS-EFL	0.009 [0.003, 0.014]	0.72	[0.37, 1.04]			
	NS 0.04 (0.01)							
ls1	EFL 0.45 (0.18)					3.54	0.030	*
	ESL 0.40 (0.18)	NS-EFL	-0.074 [-0.143, -0.005]	-0.43	[-0.77, -0.06]			
	NS 0.38 (0.16)							
ndwerz	EFL 39.33 (1.38)					3.79	0.024	*
	ESL 39.51 (1.43)	NS-EFL	0.631 [0.073, 1.189]	0.45	[0.09, 0.80]			
	NS 39.97 (1.38)							
ndwesz	EFL 37.98 (1.77)					5.62	0.004	**
	ESL 38.56 (1.58)	NS-EFL	0.947 [0.275, 1.619]	0.54	[0.21, 0.85]			
	NS 38.93 (1.70)							
logttr	EFL 0.83 (0.02)					4.28	0.015	*
	ESL 0.84 (0.03)	NS-EFL	0.011 [0.000, 0.020]	0.48	[0.15, 0.79]			
	NS 0.84 (0.03)							
uber	EFL 18.60 (1.44)	ESL-EFL	0.864 [0.047, 1.682]	0.47	[0.15, 0.75]	6.78	0.001	**
	ESL 19.46 (2.11)	NS-EFL	1.244 [0.426, 2.061]	0.61	[0.28, 0.90]			
	NS 19.84 (2.46)							

maas	EFL 0.06 (0.00)	ESL-EFL	-0.002 [-0.004, -0.000]	-0.53	[-0.85, -0.22]	9.36	<.001	***
	ESL 0.05 (0.01)	NS-EFL	-0.003 [-0.005, -0.001]	-0.71	[-1.04, -0.36]			
	NS 0.05 (0.01)							
mattr	EFL 0.74 (0.02)	ESL-EFL	0.010 [0.000, 0.021]	0.40	[0.04, 0.72]	12.28	<.001	***
	ESL 0.75 (0.03)	NS-EFL	0.022 [0.011, 0.032]	0.89	[0.52, 1.21]			
	NS 0.76 (0.03)							
msttr	EFL 0.74 (0.02)	ESL-EFL	0.012 [0.001, 0.022]	0.44	[0.08, 0.77]	13.26	<.001	***
	ESL 0.75 (0.03)	NS-EFL	0.022 [0.012, 0.033]	0.92	[0.54, 1.26]			
	NS 0.76 (0.02)	NS-ESL	0.010 [0.000, 0.021]	0.40	[0.05, 0.72]			
hdd	EFL 0.79 (0.02)					4.95	0.007	**
	ESL 0.79 (0.03)	NS-EFL	0.011 [0.002, 0.020]	0.56	[0.16, 0.90]			
	NS 0.80 (0.02)							
mtld	EFL 54.26 (7.58)	ESL-EFL	4.309 [0.205, 8.414]	0.43	[0.11, 0.75]	11.86	<.001	***
	ESL 58.57 (11.70)	NS-EFL	8.468 [4.363, 12.57]	0.89	[0.55, 1.20]			
	NS 62.73 (11.09)	NS-ESL	4.158 [0.053, 8.262]	0.36	[0.04, 0.71]			
vocd	EFL 84.72 (11.85)					5.70	0.003	**
	ESL 90.71 (17.99)	NS-EFL	8.745 [2.496, 14.99]	0.60	[0.25, 0.95]			
	NS 93.47 (16.48)							

– The number of observations for all tests is 210. The degrees of freedom for all lexical analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by bold asterisks are based on the Bonferroni-corrected alpha level of 0.002.

This can be due to the fact that vv1 searches varied verbs in verb tokens whereas vv2 searches varied verbs in lexical tokens. Considering that vv1 operates on a limited search criterion, the results suggest that the ESL group has produced more varied verbs.

Finally, in the **conclusion** rhetorical sections, we see a slightly different picture: The EFL group only outperforms marginally regarding the ls1 index compared to the English L1 group, resulting in a small effect for this comparison, and outperforms both English L1 and ESL groups relatively better regarding the maas index, resulting in medium effects of the mean-difference comparisons. Similar to the previous rhetorical sections' findings, mattr, msttr and mtld measures capture the differences with medium to large effects. In this analysis, the mean difference and CI values are statistically significant for all three sets of comparisons as demonstrated in table 6.12, the most significant one being the NS-EFL comparison with an effect size of 0.9 and the CI which reach up to 1.2. A comparison of the results sections with the aggregated corpus also supports the assumption in experts' works that the mattr, msttr and mtld measures of lexical diversity are text-length and rhetorical-section independent and can capture between-group differences regarding the production of varied vocabulary with large effects.

The D measure's variants of hdd and vocd record similar values for effect sizes and CIs, as well as the significance levels of ANOVA for the NS-EFL comparison set; the vocd index shows slightly larger values. Among the logarithm-based measures, logttr and uber show relatively more significant results. However, the logttr measure is only significant at the 0.01 level with a small effect for only one comparison set, while uber shows a p-value which is significant at the stricter 0.001 level with a medium effect for two comparison sets. Between the two measures which calculate the number of different words, ndwesz's result is significant at the 0.004 level for the NS-EFL comparison only with a medium effect. The analysis of Lexical density, however, results in the mean difference values being highly significant as marked by the p-value of the F statistic, albeit with a medium effect size of Cohen's d and the Hedge's confidence intervals which reach up to 1 for the NS-EFL comparison. These findings are in sharp contrast with Pietila (2015) analysis of conclusion sections of MA dissertations in linguistics and literature disciplines where none of the lexical density and diversity measures showed any between-group differences of English L1 vs L2 but lexical sophistication indices did show significant differences.

Overall, the findings of the conclusion sections of dissertations yield evidence to the lexical complexity of the texts of the English L1 group, followed closely by the ESL group especially with regard to the lexical diversity and density indices. The lexical sophistication of ls1, as with the findings of previous rhetorical sections, is shown to be the area that the EFL group is outperforming. These findings indicate that the English L1 and ESL groups generally use more varied and new vocabulary which is not necessarily among the less-frequently used indices as filtered through the frequently-used words in BAWE and BNC lists. The results also point to the similarity of the last three rhetorical sections regarding the presence of the three constructs of lexical complexity and their distinguishing powers of the texts of the three groups with similar effect sizes. Table 6.13 presents a summary of the statistically significant results of the lexical complexity measures that could capture between-group differences across the six rhetorical sections as well as in the whole corpus.

Table 6.13. Lexical measures that show between-group differences at the 0.002 alpha level only

Data/Rhetorical Sections	**Lexical Measures Significant at the 0.002 level only**
Aggregated Data	ld, vv2, mattr, msttr, mtld
Abstract	ndwerz, ndwesz, logttr, uber, maas, mattr, msttr, hdd, mtld, vocd
Introduction	ndwerz, ndwesz, mattr, msttr, mtld, vocd
Literature Review	ld, uber, lv, vv2, nv, maas
Method and Design	ld, ls1, mattr, msttr, mtld
Results and Discussion	ld, mattr, msttr, mtld
Conclusion	ld, maas, mattr, msttr, mtld

– The Bonferroni-corrected alpha level of (0.05 / 22 = 0.002) is applied.
– The between-group differences are based on the analyses of ANOVA, the post-hoc Tukey HSD, the bootstrapped confidence intervals and effect sizes in the previous tables.

As these tables indicate, lexical density is a good indicator of text complexity differences regarding the English language backgrounds, with English L1 followed by ESL groups producing more lexically dense texts in all rhetorical sections. Moreover, lexical density is a token-token ratio and therefore, is not affected by sample size. As the results of other studies suggest (e.g., in Kim, 2014) it makes a reliable index for finding proficiency differences in academic writing texts, and in this study across rhetorical sections with varying length.

The construct of lexical diversity, especially the measures based on word strings, was also confirmed to be a reliable distinguisher of the texts of the three groups regardless of the effect of rhetorical sections. The tables indicate that the three lexical diversity measures based on word strings/segments (i.e., mtld, mattr, msttr) better capture differences of academic texts across rhetorical sections with different length. McCarthy and Jarvis (2010) explain at length why msttr works well with longer texts but results in discarding parts of the text that are not included in word segments, and why mtld could be a better measure in this regard, i.e., because of the point of stabilisation that smoothens the TTR trajectory and the use of "an empirically driven textual factor size" instead of fixed segment sizes in msttr (p. 386). This difference can be seen in slightly better performance of mtld compared to msttr in capturing lexical diversity as indicated in the above tables.

The vocd measure only showed significant differences in shorter rhetorical sections of abstract and introduction which could be due to the sampling procedure which affects longer texts (i.e., the longer the text, it is less likely that the whole text is covered by sampling, see e.g., McCarthy & Jarvis, 2007). Slightly-better effect sizes of vocd compared to hdd could also be due to the sample sizes in the formulas of the two measures for random sampling which mainly affects longer texts (McCarthy, 2020, personal communication). This is because the sample size (e.g., 35 tokens and 36-50 tokens) in the formula of Vocd-D measure were set based on previous studies (e.g., speech segments, see Malvern et al., 2004) and not based on very long texts. This is while HD-D looks at every word in a text, albeit with a small sample size of 42 in its random sampling which may not be optimal for very long texts either. This is the main reason I initially discussed that one needs to compare the performance of these measures based on a more reliable measure, e.g., mtld. These results will be revisited in the answers to the research questions in 6.8.

Mixed results are obtained regarding lexical sophistication indices. Unlike lexical density and diversiy which showed larger values for English L1 followed by the ESL groups, lexical sophistication index of ls1 showed larger values for the EFL group, albeit with small to medium effects for between-group differences. These differences are pronounced in method, result, and conclusion sections. Upon manual inspection of texts, I found larger amounts of fied-specific terminology in these sections in the EFL texts that were closely linked with paraphrasing the experts' opinions, results, etc, for example in:

"**Metadiscourse**, according to Ädel (2006), is one type of **reflexivity** of language and **reflexivity** is a universal feature of language thus **metadiscourse** could be a universal feature as well. Using some **interactive markers** such as code **glosses**, **transitions**, and frame **markers** make the texts clear and comprehensible to the audience through minimizing the readers' processing efforts. Undoubtedly, if there are enough **transitions** and frame **markers** in a political figure's long impromptu speech, it will tell us that he or she has an arranged mind."

as well as the frequent use of the names and descriptions of various statistics (e.g., 'non-parametric Kruskal Wallis', 'asymptotic significant level', 'skewness ratio'). The analysis of a sample EFL text from the introduction texts in 6.3.2 below, on the other hand, shows that most of sophisticated lexical items are not field-specific. Regarding the ls2 index of lexical sophistication, the English L1 and ESL groups received larger values in the introduction section (see table 6.8) with small effects for the between-group differences with the EFL group. This seemingly contradictory finding regarding ls1 and ls2 can be attributed to the quantification methods of these measures in that ls1 searches for all sophisticated lexical items which are not necessarily unique while ls2 searches for sophisticated lexical types (non-repetitious). As such, ls2 has a stricter criterion and can be regarded as a hybrid measure of sophistication and diversity. The results, therefore, show similar performances of the three groups concerning the production of lexically sophisticated texts overall. This result differs from Paquot (2019) in which the values of all lexical (and verb) sophistication indices were larger in the highest proficiency group among EFL learners, that could be attributed to the type of texts (research papers vs. the entire dissertations) as well as the groups of students.

6.3.2. Some Linguistic Examples of Lexical Complexity from the Texts of the Three Groups

Since this study is mainly a quantitative analysis of rhetorical sections of MA dissertations and a multi-layered measure-testing process, a detailed qualitative analysis of the dissertations is beyond the scope of this study. However, in the following paragraphs, I include excerpts from the dissertations of the three groups as linguistic examples to discuss lexical complexity constructs and measures in context. I identified the dissertations in each group that obtained the mean values or very close to mean values for all or most of the lexical complexity measures across the rhetorical sections. I then selected 200-203 consecutive tokens (that form a complete paragraph) from the introduction sections of three of these dissertations, one in each group, for a linguistic analysis of the texts. The excerpts have similar functions as they are all part of/explain the rationale for the study and contain the wording of the students, i.e., without (extensive) citations, numbers, date, examples, etc. These excerpts are all from dissertations in the field of TEFL and SLA. I have manually identified the sophisticated lexical types in bold font. The underlined words have a similar construction to them (e.g., from another word class/part of speech) in the BNC frequency word list. The quantitative results of a handful of measures based on lemmatised texts are also included as additional information.

EFL Excerpt:

It is believed when an **eager FL** student faces an FL teacher having a great **command** on pronunciation and speaking skills different from the others, he may be encouraged to speak like him because it sounds appealing. There might be more student **preparation** for a class like that, and learning would be **facilitated indirectly** by teacher's correct pronunciation. That can be among the direct and/or indirect benefits of pronunciation **reflection** in teaching-learning process. Of course, other features of a good English language teacher, such as the ability to transfer the knowledge, should not be ignored. In countries such as **Iran**, English language is taught as a Foreign Language (FL). In cases like that, the effect of the teachers' accent on the students' seems to be something **inevitable** because of the limited students' **exposure** to English language. To study the English language pronunciation status of **Iranian EFL** teachers, getting to know some features of **Persian** pronunciation system seems to be a need. When EFL learners say 'tree' instead of 'three', they should not expect the **native** listener to get what they have wanted to produce at the first step because the addressee does not live on their mind.

Tokens= 200 Types= 115 No. sophisticated types=12 Lexical variation/lv= 0.67

Lexical sophistication/LS1= 0.33 Lexical density= 0.03 Verb variation/cvv1= 0

ESL Excerpt:

Collocations form an **integral** part of any discourse, written or spoken. However, the **partially restricted** nature of collocations makes them very challenging for second (**ESL**) and foreign language learners (**EFL**), even at **advanced** levels of **proficiency**. ESL/EFL students' **inadequate** knowledge of collocations usually **affects**, not only their comprehension of the language, but also their language production. This study will concentrate on **discursive**/ **argumentative** writing. This type of writing is selected because, **unlike** other **registers** such as **creative** writing, this register demands features like **clarity**, **precision** and lack of ambiguity . These features are preferred for the purpose of this study, as the focus is on the use of collocations and not any other **stylistics** feature. **Besides**, since the **ultimate** goal of this study is to suggest strategies for improving the learners' lexical proficiency in English, it is reasonable to focus on a register which is required from learners at advanced levels, and which can help them become more successful in their higher studies. Failing to use **native-like** expressions can create an impression of **brusqueness**. A common **limitation** of previous studies on collocations among **Arab** learners is the use of **elicitation** tests as the only tool to assess learners' knowledge of collocations.

Tokens= 203 Types= 116 No. sophisticated types= 25 Lexical Variation/lv= 0.88

Lexical sophistication/LS1= 0.25 Lexical density= 0.04 Verb variation/cvv1= 0.71

English L1 Excerpt:

Textbooks are **manuals** of instruction, typically **composed** of various organized units of work, used to **educate** students on a collection of knowledge, principles, and concepts surrounding a particular **topic** or subject. **Habitually**, textbooks are written by experts in the field, composed by publishing houses (**editorials**) to be **distributed** to schools, universities or libraries. However, textbooks are not **didactically** perfect. They are **constructed** and **simplified** in order to **convey** information to students at different academic levels. This information has to be **restrictive** as textbooks are **finite** spaces and need to be **selective** for purposes of **clarity**. This is **problematic** when considering the **protagonist** role of the textbook in the classroom . One of the first official definitions of the textbook in Spain is offered by the Instrucción Pública. This first definition is **solid** and **satisfactory**, as it emphasises not only the **necessity** for clarity and **exactitude**, but for **objectivity**, vis-à-vis reflecting current scientific knowledge. This **evidently** implies that textbooks should not be **convoluted**, complex, and/or based on **myths**, **unfounded** beliefs or **factually** incorrect information. A textbook that **possesses** such qualities is not only **counterproductive** but **contrary** to the **essence** of the textbook itself. This **idealised summary** of the textbook is too **optimistic**.

Tokens= 203 Types= 119 No. sophisticated types= 36 Lexical variation/lv= 0.77

Lexical sophistication/ls1= 0.54 Lexical density= 0.06 Verb variation/cvv1= 1

One can tell at a glance at the three excerpts that the English L1 text is demonstrably more lexically sophisticated compared to the other two texts. Moreover, these sophisticated lexical types seem to be proportionately dispersed in the English L1 text compared to the other two excerpts. Two of the sophisticated types in the EFL and ESL texts are acronyms, that are fairly common in TEFL and SLA sub-disciplines. However, these acronyms (e.g., EF, EFL, ESL) are not among the frequently-used words in the BNC nor the BAWE word lists, and therefore, classed as advanced. Their frequent equivalents in these word lists were L1 and L2. The other three lexical items of ‘Iran’, ‘Iranian’, and ‘Persian’ have also been identified as sophisticated purely because of their absence in the mentioned word lists, but they are clearly not ‘advanced’ for an Iranian EFL student. Therefore, based on a qualitative analysis of the EFL excerpt I do not label this excerpt as a sophisticated academic text in terms of its lexis, but rather as a general argumentative essay. The ESL and English L1 texts, on the other hand, seem to comply with the academic writing in general, both structurally and at the level of the individual lexical items. These points are further supported by the values of the few lexical measures. The ESL and the English L1 texts, for instance, are more lexically varied (lv values) than the EFL excerpt. This can be readily observed in the EFL text in the repeated use of general and less-frequently-used words, e.g., ‘like’, ‘seems’, and function words. The lack of specificity in the EFL text is also quite easily noticeable, e.g., the use of words and constructions such as ‘something’, ‘it is believed’, ‘of course’. The ESL excerpt, among the three texts, contains more lexically diverse types as calculated via T_{lex} / N_{lex}. When it comes to verb variation though, it is the English L1 student who outperformed the other groups, e.g., in the values of cvv1 that reflect the number of verb types. This can be seen in the use of verbs such as ‘composed’, ‘written’, and ‘constructed’ to refer to the same concept about the textbooks, as well as the verbs such as ‘convey’, ‘reflect, and ‘imply’, etc. Sophisticated nouns in the English L1 and ESL excerpts are also distinct, e.g., in the use of words such as ‘exactitude’ and‘ brusqueness’. But perhaps the most distinct aspect of the English L1 text is the use of varied and sophisticated adjectives and adverbs, such as ‘idealised’, ‘unfounded’, ‘counterproductive’, ‘habitually’, and ‘didactically’. There is a strong presence of adjectives among the sophisticated types (identified in bold font) in the English L1 text, and a strong presence of nouns as sophisticated types in the ESL text.

Finally, the values of lexical density linearly increase from the EFL to the ESL, and the English L1 texts, reflecting the ratio of the number of lexical tokens to all tokens. Overall, the English L1 student’s text is more lexically complex as indicated by the three constructs of density, diversity, and sophistication, and the EFL excerpt is the least-lexically complex one. Even though these excerpts cannot be taken in isolation when considering the overall complexity of the texts of each group, these sample texts could demonstrate, at a local level, a systems view of lexical complexity, e.g., by taking these excerpts as small-scale systems whereby a lexically dense text with non-repetitious and advanced words (dispersed proportionately across the text) can be viewed as a more complex system.

Regarding the rhetorical functions and communicative purposes of these texts and/or sentences based on the revised CARS models as examined in Lu et al., 2020 and the occurrence of lexical complexity structures, one notices the absence of sophisticated words in the move ‘establishing a research territory’ and its second step ‘real-world contextualisation’ in the EFL excerpt from the introduction section:

When EFL learners say ‘tree’ instead of ‘three’, they should not expect the **native** listener to get what they have wanted to produce at the first step because the addressee does not live on their mind.

As will be further discussed in chapter seven, the presence of such sentences with underused discipline-specific vocabulary (e.g., mispronunciation, voiced vs. voiceless sounds, digraphs [th], phonetic difficulties, misinterpretation by English L1, etc) contributes to overall lower quality of EFL texts and a lower lexically-sophisticated text. This is in contrast to the following sample sentence that is taken from the English L1 excerpt (the same move and step of real-world contextualisation) regarding the presence of more diverse and sophisticated vocabulary:

This first definition is **solid** and **satisfactory**, as it emphasises not only the **necessity** for clarity and **exactitude**, but for **objectivity**, vis-à-vis reflecting current scientific knowledge.

The difference in lexical diversity and sophistication of the EFL and the ESL groups in this study can also be seen in the following sample sentence (with a similar rhetorical move and step) that is taken from the ESL excerpt:

ESL/EFL students’ **inadequate** knowledge of collocations usually **affects**, not only their comprehension of the language, but also their language production.

Although a detailed qualitative analysis of all texts regarding the presence/absence of certain linguistic complexity features in various rhetorical functions and moves is beyond the scope of this study, these sample excerpts and sentences give a glimpse of how the underuse of certain words and structures could lead to less effective communication of ideas and overall lower quality of academic texts, e.g., as judged qualitatively.

6.3.3. Syntactic Complexity in Six Rhetorical Sections of Dissertations

Tables 6.14 to 6.20 follow the same process for the syntactic indices. The significant differences based on the new Bonferroni-corrected alpha level of 0.004 for the syntactic dataset are indicated by bold asterisks. The specification of column names and tests and the interpretation of the effect sizes are the same as the lexical analyses in 6.3.1. Since previous studies have not investigated these measures in various rhetorical sections/sub-genres of academic writing, I will only discuss the findings of this study based on the results of each table and include some brief discussions of any similar work at the end of this section.

Table. 6.14. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the abstract sections on the syntactic dataset

Abstract		**Tukey HSD group differences**				**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 22.04 (4.95)	ESL-EFL	3.057 [0.645, 5.469]	0.51	[0.20, 0.79]	5.19	0.006	**
	ESL 25.10 (6.76)	NS-EFL	2.586 [0.174, 4.998]	0.45	[0.12, 0.76]			
	NS 24.62 (6.28)							
VP/T	EFL 2.19 (0.52)	ESL-EFL	0.300 [0.044, 0.556]	0.49	[0.16, 0.79]	5.13	0.006	**
	ESL 2.49 (0.68)	NS-EFL	0.300 [0.045, 0.556]	0.48	[0.14, 0.77]			
	NS 2.49 (2.71)							
C/T	EFL 1.56 (0.26)	ESL-EFL	0.167 [0.031, 0.302]	0.51	[0.19, 0.79]	9.18	<.001	***
	ESL 1.73 (0.38)	NS-EFL	0.240 [0.104, 0.375]	0.75	[0.41, 1.04]			
	NS 1.80 (0.37)							
DC/C	EFL 0.29 (0.11)	ESL-EFL	0.057 [0.010, 0.105]	0.51	[0.15, 0.81]	9.60	<.001	***
	ESL 0.35 (0.12)	NS-EFL	0.086 [0.038, 0.133]	0.71	[0.38, 1.03]			
	NS 0.38 (0.13)							
DC/T	EFL 0.48 (0.23)	ESL-EFL	0.159 [0.028, 0.290]	0.52	[0.20, 0.78]	10.01	<.001	***
	ESL 0.64 (0.36)	NS-EFL	0.244 [0.113, 0.375]	0.78	[0.46, 1.08]			
	NS 0.72 (0.38)							
CT/T	EFL 0.38 (0.16)	ESL-EFL	0.096 [0.019, 0.174]	0.55	[0.19, 0.86]	11.77	<.001	***
	ESL 0.47 (0.19)	NS-EFL	0.157 [0.080, 0.234]	0.79	[0.45, 1.11]			
	NS 0.53 (0.23)							

CN/T	EFL 3.23 (0.84)	ESL-EFL	0.441 [0.025, 0.857]	0.44	[0.13, 0.75]	4.68	0.010	*
	ESL 3.67 (1.13)	NS-EFL	0.489 [0.073, 0.905]	0.49	[0.17, 0.79]			
	NS 3.72 (1.13)							

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

Table. 6.15. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the introduction sections on the syntactic dataset

Introduction		**Tukey HSD group differences**				**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 21.34 (3.37)	ESL-EFL	3.326 [1.520, 5.132]	0.73	[0.40, 1.01]	9.96	<.001	***
	ESL 24.66 (5.49)	NS-EFL	2.333 [0.527, 4.139]	0.59	[0.23, 0.87]			
	NS 23.67 (4.47)							
MLC	EFL 12.82 (1.80)	ESL-EFL	1.261 [0.377, 2.145]	0.56	[0.23, 0.86]	6.49	0.001	**
	ESL 14.09 (2.61)							
	NS 13.04 (2.17)	NS-ESL	-1.045 [-1.929, -0.161]	-0.43	[-0.74, -0.10]			
C/T	EFL 1.67 (0.19)					6.55	0.001	**
	ESL 1.77 (0.30)	NS-EFL	0.158 [0.054, 0.262]	0.66	[0.31, 0.96]			
	NS 1.83 (0.28)							

DC/C	EFL 0.33 (0.06)	ESL-EFL	0.038 [0.005, 0.070]	0.46	[0.14, 0.82]	7.62	<.001	***
	ESL 0.37 (0.10)	NS-EFL	0.051 [0.019, 0.084]	0.73	[0.40, 1.08]			
	NS 0.39 (0.08)							
DC/T	EFL 0.57 (0.15)	ESL-EFL	0.118 [0.023, 0.212]	0.51	[0.17, 0.82]	8.45	<.001	***
	ESL 0.68 (0.29)	NS-EFL	0.158 [0.063, 0.252]	0.76	[0.42, 1.05]			
	NS 0.72 (0.250)							
CT/T	EFL 0.40 (0.10)	ESL-EFL	0.060 [0.011, 0.110]	0.49	[0.15, 0.82]	15.11	<.001	***
	ESL 0.46 (0.14)	NS-EFL	0.115 [0.066, 0.165]	1	[0.65, 1.31]			
	NS 0.51 (0.13)	NS-ESL	0.054 [0.005, 0.104]	0.40	[0.08, 0.73]			
CN/T	EFL 3.11 (0.62)	ESL-EFL	0.354 [0.039, 0.670]	0.46	[0.12, 0.76]	3.75	0.024	*
	ESL 3.46 (0.89)							
	NS 3.37 (0.83)							

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

Considering the syntactic analysis in the **abstract** sections (table 6.14), out of eleven total measures, four measures of C/T (clauses per T-unit), DC/C (dependent clauses per clause), DC/T (dependent clauses per T-unit), and CT/T (complex T-units per T-unit) showed statistically significant mean-difference values with medium effects for the ESL-EFL and NS-EFL comparison sets with the confidence intervals of effect sizes that reach above 1 for all four comparisons of NS-EFL. The other three measures of MLT (mean length of T-unit), VP/T (verb phrases per T-unit) and CN/T (complex nominals per T-unit) did not show any significant mean differences. A quick look at table 6.14 indicates similar performances of the English L1 and ESL groups; we also notice the EFL group's underuse of complex syntactic structures.

As for the important constructs in this section, apart from coordination, the rest of syntactic constructs of mean length of production units, subordination, and phrasal complexity were shown to be effective in capturing text differences of postgraduate academic writing by students with different English language backgrounds. The dominant construct, however, is subordination with all four representative indices that obtained statistically significant results indicating greater amount of subordination in the abstracts of both English L1 and ESL groups.

Similar patterns of the type of syntactic indices and the group comparisons appear in the findings of the **introduction** section (table 6.15) as well. The six measures of MLT, MLC, C/T, DC/C, DC/T, and CT/T recorded statistically significant between-group differences which pass the criterion of the new alpha level set by the Bonferroni correction. The significance is further confirmed with medium effects and the Hedge's g confidence intervals as large as 1 for the NS-EFL comparisons. The only unusual result was the mean difference values of MLC (mean length of clause): the ESL group wrote longer clauses. The findings of the other five indices indicate that the English L1 group produced the most and the EFL group produced the least syntactically-complex introductions.

Regarding the important constructs in distinguishing the texts of the three groups, only the two constructs of subordination and length of production units could capture group differences with very similar effect sizes overall. Complex nominals only marginally showed some differences between the ESL and EFL groups, suggesting that all three groups have produced similar amounts of complex nominals and verb phrases in the introduction sections. These results will be revisited in the predictive modelling of important indices to classify the introduction section in 6.7.2.

Table. 6.16. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the literature review sections on the syntactic dataset

Lit. Review			**Tukey HSD group differences**			**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 21.07 (2.37)	ESL-EFL	2.912 [1.530, 4.295]	0.81	[0.46, 1.11]	12.48	<.001	***
	ESL 23.98 (4.44)	NS-EFL	1.697 [0.315, 3.080]	0.59	[0.25, 0.91]			
	NS 22.76 (3.27)							
MLC	EFL 11.92 (1.05)	ESL-EFL	0.614 [0.012, 1.216]	0.37	[0.04, 0.67]	3.19	0.043	*
	ESL 12.53 (2.05)							
	NS 12.06 (1.24)							
VP/T	EFL 2.39 (0.26)	ESL-EFL	0.151 [0.003, 0.300]	0.41	[0.10, 0.70]	3.26	0.040	*
	ESL 2.55 (0.44)							
	NS 2.52 (0.39)							
C/T	EFL 1.77 (0.18)	ESL-EFL	0.153 [0.056, 0.250]	0.64	[0.31, 0.93]	7.77	<.001	***
	ESL 1.93 (0.29)	NS-EFL	0.122 [0.025, 0.219]	0.55	[0.22, 0.87]			
	NS 1.89 (0.250							
DC/C	EFL 0.38 (0.05)	ESL-EFL	0.031 [0.006, 0.055]	0.51	[0.19, 0.86]	5.83	0.003	**
	ESL 0.41 (0.07)	NS-EFL	0.030 [0.005, 0.054]	0.53	[0.19, 0.88]			
	NS 0.41 (0.06)							
DC/T	EFL 0.68 (0.15)	ESL-EFL	0.129 [0.042, 0.216]	0.60	[0.30, 0.90]	7.27	<.001	***
	ESL 0.81 (0.26)	NS-EFL	0.111 [0.024, 0.198]	0.57	[0.25, 0.90]			
	NS 0.80 (0.23)							

CT/T	EFL 0.47 (0.07) ESL 0.53 (0.10) NS 0.53 (0.10)	ESL-EFL NS-EFL	0.064 [0.027, 0.101] 0.065 [0.027, 0.102]	0.73 0.73	[0.37, 1.07] [0.41, 1.06]	11.21	<.001	***
CP/C	EFL 0.46 (0.11) ESL 0.42 (0.16) NS 0.40 (0.11)	NS-EFL	-0.063 [-0.116, -0.010]	-0.55	[-0.86, -0.23]	4.33	0.014	*
CN/T	EFL 2.98 (0.37) ESL 3.43 (0.81) NS 3.22 (0.60)	ESL-EFL NS-EFL	0.455 [0.208, 0.702] 0.247 [0.000, 0.494]	0.72 0.49	[0.42, 1.02] [0.15, 0.79]	9.48	<.001	***

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

Table. 6.17. Between-group differences, ANOVA, and post-hoc effect siezes and confidence intervals for the method & design sections on the syntactic dataset

Method			**Tukey HSD group differences**			**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 19.69 (2.80) ESL 22.58 (3.58) NS 21.97 (3.01)	ESL-EFL NS-EFL	2.887 [1.632, 4.143] 2.272 [1.016, 3.528]	0.89 0.78	[0.56, 1.19] [0.40, 1.11]	16.35	<.001	***

VP/T	EFL 2.07 (0.28)	ESL-EFL	0.319 [0.188, 0.451]	0.95	[0.55, 1.26]	21.84	<.001	***
	ESL 2.39 (0.39)	NS-EFL	0.317 [0.186, 0.449]	1.06	[0.64, 1.39]			
	NS 2.38 (0.31)							
C/T	EFL 1.58 (0.17)	ESL-EFL	0.189 [0.103, 0.275]	0.90	[0.51, 1.22]	17.45	<.001	***
	ESL 1.77 (0.24)	NS-EFL	0.182 [0.096, 0.268]	0.91	[0.55, 1.22]			
	NS 1.76 (0.22)							
DC/C	EFL 0.29 (0.06)	ESL-EFL	0.071 [0.044, 0.0992]	1	[0.64, 1.33]	27.52	<.001	***
	ESL 0.36 (0.08)	NS-EFL	0.077 [0.049, 0.104]	1.23	[0.87, 1.53]			
	NS 0.37 (0.06)							
DC/T	EFL 0.46 (0.14)	ESL-EFL	0.190 [0.114, 0.265]	1.01	[0.65, 1.31]	23.72	<.001	***
	ESL 0.65 (0.22)	NS-EFL	0.191 [0.116, 0.267]	1.13	[0.79, 1.43]			
	NS 0.66 (0.19)							
CT/T	EFL 0.34 (0.09)	ESL-EFL	0.110 [0.068, 0.151]	1.01	[0.66, 1.36]	28.2	<.001	***
	ESL 0.45 (0.13)	NS-EFL	0.119 [0.077, 0.161]	1.30	[0.92, 1.64]			
	NS 0.46 (0.10)							
CN/T	EFL 2.62 (0.47)	ESL-EFL	0.32 [0.104, 0.538]	0.57	[0.23, 0.87]	6.18	0.002	**
	ESL 2.94 (0.65)							
	NS 2.81 (0.50)							

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.

– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

Table 6.16 demonstrates rather similar patterns with regard to the type and number of syntactic indices which captured between-group differences in the l**iterature review** sections: the six measures of MLT (mean length of T-unit), C/T (clauses per T-unit), DC/C (dependent clauses per clause), DC/T (dependent clauses per T-unit), CT/T (complex T-units per T-unit), and CN/T (complex nominals per T-unit) once more record medium effects for the ESL-EFL as well as NS-EFL comparisons. Furthermore, the p-values of the analyses of variance well meet the strict requirement of the alpha level set by the Bonferroni-correction method. The ESL-EFL comparison set, however, resulted in larger values of the point estimate effect size and their corresponding confidence intervals. This finding, together with the mean values of the three groups, suggest that the ESL group produced slightly more complex syntactic structures than the other groups regarding the mentioned syntactic indices. The only measure in this section which distinguishes the EFL group's outperformance is the CP/C (coordinate phrases per clause) index which resulted in a medium effect for the NS-EFL comparison that is significant only at the 0.01 level. This finding is quite similar to the aggregated syntactic data. Overall, the most syntactically-complex literature review texts are produced by the ESL group, followed closely by English L1s.

Considering useful constructs for distinguishing syntactic differences of the three groups' texts, the three constructs of subordination, length of production units, and phrasal complexity were found effective with similar effects. For the first time, we observe that the number of complex nominal structures is significantly larger in the English L1 group, followed by the ESL group. The findings suggest that very long sections of literature reviews may elicit greater amounts of phrasal structures and longer and more complex T-units compared to other structures.

The analysis of the **method and design** section as presented in table 6.17 reveals the exact same syntactic measures indicating the between-group differences as the conclusion section that will be discussed afterwards. Similar results are also spotted in the aggregated data that is presented in table 6.4. In the method sections, the seven indices of MLT, VP/T, C/T, DC/C, DC/T, CT/T, and CN/T all specify the values of between-group mean differences and their corresponding CIs with medium to large effect sizes which record as large as 1.30 for the NS-EFL comparison. Furthermore, the F statistic p-values for all these measures fulfil the stringent assumption of the new Bonferroni-corrected alpha level. A quick eyeballing of the mean values of the three groups in the mentioned indices also indicates that the ESL and English L1 groups performed very similar, and both groups outperformed the EFL group in the production of more syntactically-complex texts, especially more complex T-units.

Method sections have been so far the most syntactically-complex sections in terms of a stronger presence of the three constructs of ‘length of production units’, ‘subordination’, and ‘phrasal complexity’ as distinguishers of the three groups’ texts. The ESL group has also produced the most syntactically complex method sections among the groups, as indicated by larger values of most of the indices. This is in contrast to the results of lexically complex method sections where the English L1 and EFL groups both produced more diverse and sophisticated texts overall. The results of very low correlations between lexical and syntactic measures in section 6.4 confirms that complex syntactic structures do not necessarily elicit more diverse lexical items within those structures.

For the first time, the CP/C (coordinate phrases per clause) index records statistically significant results for the EFL group in the **result and discussion** rhetorical section (table 6.18). This measure has already distinguished the EFL group’s performance in the literature review section at the 0.01 significance level for the NS-EFL comparison only; in the result section, however, this distinction is marked at the .<000 level for both ESL-EFL and NS-EFL comparisons. The increased use of coordination as opposed to subordination is believed to be a characteristic of less syntactically-proficient students and that the progression follows the pattern of coordination to subordination and then to phrasal elaboration (see for instance the discussion in Kuiken & Vedder, 2019). Therefore, larger values of CP/C and CP/T (coordinate phrases per T-unit) indices compared to other measures for the EFL group may indeed support the assumption that they are less syntactically proficient.

The rest of indices which indicated significant between-group mean differences in the result sections are MLT, VP/T, C/T, DC/C, DC/T, CT/T and CN/C. The mean and mean difference values in these sections denote the syntactic proficiency of the ESL group over the other two groups as indicated by the effect sizes, CIs and p-values. The ESL students outperformed the English L1s regarding the values of the five structures of mean length of T-units, verb phrases, complex T-units, clauses per T-units, and complex nominals. This seems to be the only rhetorical section in which the ESL group dominantly produced larger amounts of syntactically complex structures, followed by the method section as discussed before.

The result sections, compared to the previous sections, witnessed the dominance of subordination in distinguishing group performances, especially regarding dependent clauses and complex T-units per T-units for the ESL-EFL comparison set. Even though English L1s produced greater amounts of phrasal complexity than the other groups, the differences are not as noticeable as subordination indices.

Table. 6.18. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the results & discussion sections on the syntactic dataset

Results				**Tukey HSD group differences**			**ANOVA**	
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen’s d**	**Bootstrapped Effect size [95% BCa CIs] Hedges’ g**	**F**	**Pr(>F)**	**Sig.**
MLT								
	EFL 21.39 (3.14)	ESL-EFL	2.281 [0.774, 3.789]	0.61	[0.28, 0.94]	8.33	<.001	***
	ESL 23.67 (4.18)							
	NS 21.44 (3.94)	NS-ESL	-2.234 [-3.741, -0.726]	-0.55	[-0.87, -0.18]			
VP/T	EFL 2.31 (0.40)	ESL-EFL	0.331 [0.152, 0.511]	0.80	[0.43, 1.13]	10.14	<.001	***
	ESL 2.64 (0.43)							
	NS 2.40 (0.51)	NS-ESL	-0.237 [-0.417, -0.058]	-0.50	[-0.85, -0.13]			
C/T	EFL 1.75 (0.18)	ESL-EFL	0.228 [0.122, 0.335]	0.97	[0.65, 1.28]	12.8	<.001	***
	ESL 1.98 (0.28)	NS-EFL	0.108 [0.002, 0.215]	0.41	[0.08, 0.69]			
	NS 1.86 (0.32)	NS-ESL	-0.119 [-0.226, -0.013]	-0.39	[-0.74, -0.05]			
DC/C	EFL 0.36 (0.06)	ESL-EFL	0.062 [0.034, 0.090]	0.95	[0.60, 1.26]	14.21	<.001	***
	ESL 0.42 (0.07)	NS-EFL	0.036 [0.008, 0.064]	0.51	[0.20, 0.84]			
	NS 0.40 (0.08)							
DC/T	EFL 0.64 (0.16)	ESL-EFL	0.213 [0.118, 0.309]	1.01	[0.67, 1.29]	13.99	<.001	***
	ESL 0.86 (0.25)	NS-EFL	0.121 [0.025, 0.216]	0.51	[0.22, 0.78]			
	NS 0.76 (0.29)							
CT/T	EFL 0.44 (0.09)	ESL-EFL	0.094 [0.053, 0.134]	1.03	[0.67, 1.38]	15.17	<.001	***
	ESL 0.54 (0.09)	NS-EFL	0.052 [0.011, 0.092]	0.49	[0.13, 0.80]			

	NS 0.50 (0.12)	NS-ESL	-0.041 [-0.082, -0.001]	-0.39	[-0.71, -0.04]			
CP/T	EFL 0.64 (0.16)					3.72	0.025	*
	ESL 0.61 90.18)	NS-EFL	-0.079 [-0.148, -0.009]	-0.46	[-0.81, -0.09]			
	NS 0.56 (0.19)							
CP/C	EFL 0.37 (0.08)	ESL-EFL	-0.054 [-0.089, -0.019]	-0.66	[-0.99, -0.29]	10.48	<.001	***
	ESL 0.31 (0.08)	NS-EFL	-0.061 [-0.095, -0.026]	-0.69	[-1.05, -0.29]			
	NS 0.30 (0.09)							
CN/T	EFL 3.08 (0.54)					5.03	0.007	**
	ESL 3.25 (0.72)							
	NS 2.91 (0.64)	NS-ESL	-0.341 [-0.595, -0.087]	-0.50	[-0.85, -0.16]			
CN/C	EFL 1.75 (0.25)					5.99	0.002	**
	ESL 1.65 (0.32)	NS-EFL	-0.176 [-0.298, -0.055]	-0.59	[-0.95, -0.23]			
	NS 1.58 (0.34)							

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

Table. 6.19. Between-group differences, ANOVA, and post-hoc effect sizes and confidence intervals for the conclusion sections on the syntactic dataset

Conclusion		**Tukey HSD group differences**				**ANOVA**		
Syntactic Measures	**Mean and (SD)**	**Group Comparisons**	**Mean difference & [95% BCa CIs]**	**Effect size Cohen's d**	**Bootstrapped Effect size [95% BCa CIs] Hedges' g**	**F**	**Pr(>F)**	**Sig.**
MLT	EFL 22.45 (3.71)	ESL-EFL	2.905 [1.022, 4.788]	0.62	[0.29, 0.93]	7.68	<.001	***
	ESL 25.36 (5.49)	NS-EFL	2.453 [0.570, 4.336]	0.57	[0.22, 0.85]			
	NS 24.91 (4.78)							
VP/T	EFL 2.42 (0.36)	ESL-EFL	0.353 [0.145, 0.561]	0.70	[0.35, 0.97]	11.7	<.001	***
	ESL 2.77 (0.62)	NS-EFL	0.382 [0.174, 0.589]	0.82	[0.46, 1.12]			
	NS 2.80 (0.55)							
C/T	EFL 1.74 (0.22)	ESL-EFL	0.183 [0.071, 0.296]	0.68	[0.34, 1.01]	16.92	<.001	***
	ESL 1.93 (0.31)	NS-EFL	0.272 [0.159, 0.385]	1.01	[0.65, 1.35]			
	NS 2.02 (0.31)							
DC/C	EFL 0.36 (0.07)	ESL-EFL	0.062 [0.030, 0.093]	0.78	[0.38, 1.13]	18.83	<.001	***
	ESL 0.42 (0.09)	NS-EFL	0.076 [0.045, 0.108]	1.05	[0.69, 1.37]			
	NS 0.44 (0.08)							
DC/T	EFL 0.64 (0.19)	ESL-EFL	0.198 [0.094, 0.301]	0.79	[0.49, 1.13]	19.23	<.001	***
	ESL 0.84 (0.29)	NS-EFL	0.260 [0.156, 0.363]	1.08	[0.71, 1.42]			
	NS 0.90 (0.28)							
CT/T	EFL 0.45 (0.11)	ESL-EFL	0.079 [0.032, 0.126]	0.66	[0.30, 0.99]	21.3	<.001	***
	ESL 0.53 (0.13)	NS-EFL	0.128 [0.081, 0.175]	1.18	[0.77, 1.53]			
	NS 0.58 (0.11)	NS-ESL	0.048 [0.001, 0.095]	0.39	[0.06, 0.72]			

CN/C	EFL 1.92 (0.36)					3.14	0.045	*
	ESL 1.88 (0.47)	NS-EFL	-0.162 [-0.321, -0.003]	-0.45	[-0.79, -0.13]			
	NS 1.76 (0.35)							

– The number of observations for all tests is 210. The degrees of freedom for all syntactic analyses of variance are 2 and 207.
– The significant results of ANOVA as shown by asterisks are based on the new Bonferroni-corrected alpha level of 0.004.

As elaborated in the discussion of the method and design section, the type of measures and group comparisons with significant results in the **conclusion** section are almost identical to the method section. In this section also the six measures of MLT, VP/T, C/T, DC/C, DC/T, and CT/T consistently captured statistically significant results for the ESL-EFL and NS-EFL comparisons with medium to large effects that reach up to 1.18 for the NS-EFL comparison of the CT/T index. The EFL group only marginally outperforms the English L1 group regarding the CN/C (complex nominals per clause) index with a small effect. Considering the rest of indices, English L1s followed closely by the ESL group produced rather similar amounts of complex syntactic structures.

The conclusion sections are also among the three most-syntactically-complex rhetorical sections (alongside the methods and result sections) both in terms of the greater amounts of complex syntactic structures produced by the three groups overall, and in terms of the distinguishing powers of the three constructs of length of production units, subordination, and phrasal complexity in capturing text differences that include the EFL group. All three rhetorical sections, for instance, show the strong presence of verb phrases mainly in the English L1 and ESL texts and greater values of VP/T for differences between these groups and EFLs.

Table 6.20. Syntactic measures that show between-group differences at the 0.001 alpha level only

Data/Rhetorical Sections	**Syntactic Measures Significant at the 0.001 level only**
Aggregated Data	MLT, C/T, CT/T, DC/C, DC/T, VP/T
Abstract	C/T, DC/C, DC/T, CT/T
Introduction	MLT, MLC, C/T, DC/C, DC/T, CT/T
Literature Review	MLT, C/T, DC/C, DC/T, CT/T, CN/T
Method and Design	MLT, VP/T, C/T, DC/C, DC/T, CT/T, CN/T
Results and Discussion	MLT, VP/T, C/T, DC/C, DC/T, CT/T, CN/C, CP/C
Conclusion	MLT, VP/T, C/T, DC/C, DC/T, CT/T

– The between-group differences are based on the ANOVA, the post-hoc Tukey HSD, the bootstrapped confidence intervals and effect sizes.

A glance at table 6.20 reveals that by and large all syntactic measures investigated in this study captured between-group differences at the Bonferroni-corrected level of 0.004, but the five indices of MLT, C/T, CT/T, DC/C, and DC/T consistently showed up in group comparisons in all rhetorical sections. This provides evidence to the reliability of these measures in finding group differences regardless of the type of texts (i.e., not being dependent on rhetorical sections) in similar research contexts and proficiency levels as this study.

The results of syntactic complexity values across various rhetorical sections corroborate the overall assumption that coordination is used more by the students at lower proficiency levels and that subordination and phrasal-level complexity are often produced more by students at higher proficiency levels (see the discussions in Ai & Lu, 2013; Bardovi-Harlig & Bofman, 1989; Crossley & McNamara, 2009; Grant & Ginther, 2000; Mancilla et al., 2015; Norris & Ortega, 2009; Paquot, 2019; Yoon, 2017). However, unlike the claims of Biber and Gray (2010, 2013, 2016), Crossley and McNamara (2014), and Bulté and Housen (2014) that phrasal complexity, e.g., the amount of nominalisation is the best indicator of (advanced) academic writing, in this study subordination indices showed larger effect sizes for the NS-EFL comparison sets, suggesting that English L1s as the highest proficiency level in this study, produced texts with larger amounts of subordination structures than complex nominals. Liu & Li (2016) conclude that noun phrase complexity is higher in published articles by expert writers than master's dissertations (Chinese EFL in applied linguistics). This trend in complex nominals can also be seen in research papers of lower vs. higher proficiency levels of EFL learners in Paquot (2019). Furthermore, in chapter three, section 3.2 I elaborated how L2 writing researchers such as Ortega (2000) and Wolfe-Quintero et al. (1998) argue that syntactic complexity in lower proficiency levels starts with an abundance of coordination, progresses in higher proficiency levels with an abundance of subordination, and finally to more phrasal elaboration/complexity. It will be beneficial if future researchers carry out similar research to this study but with a corpus of PhD theses to find out whether English L1s produce more phrasal complexity at a more advanced level (e.g., the doctoral level) as suggested by the-mentioned researchers or that they still produce larger amounts of subordination as this study's results indicate. Paquot (2019) also argues that phraseological complexity indices could be better indicators of L2 academic writing performance at advanced proficiency levels compared to lexical and syntactic complexity measures. It is worthwhile, therefore, to conduct a similar study as this thesis and to compare the efficacy of various phraseological and collocational complexity measures to all lexical and syntactic complexity measures studied in this research.

This study’s findings also support the evidence in previous works that the length of clauses and T-units differ in English L1 vs. L2 academic writings. In Lu and Ai (2015) for instance, this difference is between the combined English L2 groups (with different L1s) vs. English L1 in the values of MLC in argumentative essays, and in Ai and Lu (2013) this difference is between Chinese EFL learners vs. English L1s regarding the MLT and MLS values. MLC values also showed an increase in research papers produced by lower vs. higher EFL proficiency levels in Paquot (2019).

Regarding proficiency differences, reaching a definite conclusion regarding the syntactic proficiency of English L1 vs. L2 writing is very difficult due to considerable variability in research designs, e.g., sample sizes, the type and length of corpora, the quantification methods especially the unit of measurements, and the effects of task types, topics, timing conditions, and the classification of proficiency levels based on holistic ratings, programme levels and external reference points like IELTS, TOEFL, and CEFR levels, as well as the English language backgrounds and students’ L1 (see the related discussions in Ai & Lu, 2013; Lu & Ai, 2015; Mancilla et al., 2015; Yoon, 2017). It seems plausible, therefore, that future researchers rely on the syntactic measures that can consistently capture proficiency differences in works with various research designs.

6.3.4. Some Linguistic Examples of Syntactic Complexity from the Texts of the Three Groups

The process of selecting excerpts from the dissertations for syntactic analyses follows that of the lexical analysis as discussed in 6.3.2 except that, these texts are selected from results & discussion sections of the dissertations from TEFL and corpus linguistics sub-disciplines. The quantitative results of a handful of production units and measures are also included as additional information. Because L2SCA counts contracted and possessive forms as two separate words, there appear to be slight differences in the number of words in these passages. Since the identification of different production units in the excerpts may override in the following texts, here I only identify/underline coordinating phrases, which is specified as adjective, adverb, noun, and verb phrases that immediately dominate a coordinating conjunction, e.g., AdjP| AdvP| NP|VP < CC (see e.g., Lu, 2010). In the ESL and English L1 excerpts, the ‘%’ symbol is taken as the alphabetic equivalent of it (‘percent’).

EFL Excerpt:

As the analysis above demonstrates, there exists no significant difference between the experimental and control group's performance on the posttests of Reading and Writing and data failed to reject the null hypotheses. And also the results agree with the research result of McClure (1990). Cooperative learning and its methods were proved to be effective through the times. But there are some limitations to implement the cooperative learning approach. The limitations to cooperative learning can be because of not structuring the cooperative learning properly and the lack of the basic elements in implementing it. If the teachers just put the students into groups to learn and didn't structure the positive interdependence and individual accountability, then it would not be unusual to find groups where one person did most(or all) of the work and the others signed off as if they had learned it or had done the work. Sometimes the students are used to being competitive or working individually and it would be hard for them to accept the cooperative learning after years being competitive. Sometimes helping low students and the existence of these so called bossy students may have some disadvantages for the class. This disadvantage and students' negative attitude, affected some researches in which the students showed negative reactions to cooperative learning. And also the limited time of this study didn't allow getting the students familiar to the concept and advantages of cooperative learning.

Words= 240   Verb phrases/VP= 32   Complex T-units= 5   Complex nominals=26

DC/C= 0.36   DC/T= 0.53   CP/T= 0.84   CN/T= 2

ESL Excerpt:

The results of the data analysis of the use of verb-noun collocations (tokens) in both corpora have revealed a statistically significant difference between advanced Arab learners and A-level native speakers. Arab learners use more verb-noun collocations tokens than native speakers. However, when correct verb-noun collocation types are compared the difference between the two corpora is not significant which shows that the frequency of correct verb noun collocation types is relatively similar in both corpora. These results contradict the findings of some elicitation studies on Arab learners which conclude that learners tend to avoid unfamiliar collocations, and the findings of Laufer and Waldman (2011), who conclude that learners at different levels of proficiency produce far fewer collocations than native speakers. It is evident in this study that Arab learners at an advanced level of proficiency did not avoid the use of verb-noun collocations but produced more verb-noun collocation tokens and a comparable number of correct verb-noun collocation types to native speakers. The type/ token ratio reveals that although advanced Arab learners use a comparable number of correct verb-noun collocation types to native speakers, the lexical diversity of these correct collocations is less as the type/token ratio is lower which indicates that learners tend to frequently repeat the collocations they use. Although Arab learners have misused adjective noun collocations less frequently than verb-noun collocations, 70.37% of the misused adjective-noun collocations are due to the influence of Arabic.

Words= 234   Verb phrases/VP= 24   Complex T-units= 4   Complex nominals=50

DC/C= 0.71 DC/T= 2.14 CP/T= 0.71 CN/T= 7.14

English L1 Excerpt:

A first significant finding, however, indicates that because Romanians and Bulgarians are largely represented concurrently in both broadsheet and tabloid newspapers, how one is represented in newspaper discourse is bound to affect the representation of the other. Quantitative analysis identified that 61.82% and 56.01% of occurrences of the lemma Romanian co-occurred within a co-text window span of 5 left/right of the lemma Bulgarian, in the broadsheet and tabloid subcorpora respectively. Considering the discursive strategies employed to contextualise the movement of migrants from Romania, it can be argued that it largely contributes to the negative representation of Romanians. The British media indicate that Romanian migration is a highly probable, large, uncontrollable, one-way phenomenon. Both newspapers also indicate a high degree of certainty in regards to the movement of migrants, achieved through the use of gerund phrases, nominalisation, as well as the salient choice of the simple present tense. The choice of tense suggests that British newspapers recontextualise a perceived phenomenon or threat into fact; this is particularly so since British newspapers are linguistically construing their own beliefs/ideologies based on their perceived impact of the transitional restrictions ending and consequent Romanian migration. Tabloids employ much more frequently direct methods of negative representation in terms of jobs. This is not necessarily surprising, and is noted as common practice by a large number of related studies. For example jobs are often represented as our or UK jobs.

Words= 235 Verb phrases/VP= 18 Complex T-units= 5 Complex nominals=38

DC/C= 0.43 DC/T= 0.77 CP/T= 0.66 CN/T= 4.22

Reading the passages, one observes the noticeable number of coordination (coordinating phrases) in the EFL excerpt (11) compared to the ESL (5) and English L1 (6) texts. Even so, one instance of such coordinations in the ESL text is a citation format (e.g., Laufer and Waldman, 2011) which might not ordinarily be regarded as an instance of phrasal coordination. This trend can also be seen in the values of CP/T that linearly decrease from EFL to ESL to English L1. As discussed in the previous chapter, Ai and Lu (2013) noticed this pattern among EFL vs English L1s' academic writing.

Features representing phrasal complexity differ in the excerpts of the three groups. The number of verb phrases, for example, is much larger in the EFL group than the other two groups: the values decrease linearly from EFL to ESL to English L1's. The number of complex nominals, on the other hand, is significantly larger in the ESL text, followed by the English L1's. This pattern is also reflected in the values of CN/T (complex nominals per T-unit). This pattern of the use of more coordination and less phrasal complexity structures, as discussed in chapter three, is believed to indicate lower English L2 proficiency levels.

Among the three categories of complex nominals as
1) noun + adjective|possessive|prepositional phrase|relative clause|participle| appositive,
2) nominal clauses, and
3) gerunds and infinitives in subject positions,
the English L1 excerpt includes more of the first pattern, e.g., the frequent pattern of ‘Noun + PP’ of ‘representation of the other’, ‘occurrences of the lemma’, ‘span of 5 left/right’, ‘movement of migrants’, ‘use of gerund’, etc. The examples of the second pattern are ‘that it largely contributes ...’, ‘that Romain migration is ...’, ‘that British newspapers recontextualise ...’, etc. The ESL excerpt also includes the first pattern more frequently, e.g., in the same ‘Noun + PP’ structure of ‘results of the data’, ‘analysis of the use’, ‘tokens in both corpora’, ‘difference between advanced’, etc followed by the second pattern, e.g., in ‘that the frequency of ...’, ‘that learners tend to ...’, ‘that Arab learners ...’, ‘elicitation studies … which conclude that ...’, etc. Similarly, the EFL excerpt demonstrates the use of first and third patterns in ‘difference between the experimental ...’, ‘performance on the posttests’, ‘reactions to cooperative learning’, and ‘groups where one person ...’, as well as in ‘helping low students ...’. The phrasal complexity patterns show that the EFL text is more verbal (verb phrases) than nominal, as opposed to the pattern found in the ESL and English L1 excerpts.

The ESL student has also produced greater amounts of subordination as reflected in the number of T-units and the values of DC/C and DC/T: the use of subordination in the ESL text is significantly higher than the other two groups. Dependent clauses in this study are specified as finite adjective, adverbial, and nominal clauses. A few examples of such constructions in the ESL text are ‘although advanced Arab learners use a comparable number of correct verb-noun collocation types to native speakers, ...’, ‘ when correct verb-noun collocation types are compared,...’, and ‘… who conclude that ...’.

Overall, and based on the discussion on the trajectory of syntactic complexification in 3.2, it seems that the ESL excerpt is the most-syntactically-complex text among the three texts, followed by the English L1 text. The EFL passage is distinctly coordinate and verbal (verb phrases) in structure followed by moderate amounts of subordination.

With regard to the rhetorical functions of syntactically complex sentences, Lu et al. (2020) examined a large-scale corpus of research articles by expert writers based on the revised CARS model and found that the move ‘presenting the present work’ and its step ‘announcing and discussing results’ is associated with higher amounts of finite dependent clauses. This association is more noticeable in the ESL and English L1 sample texts presented here. The English L1 text, for instance, has a total of eight finite dependent clauses in one paragraph, e.g., in ‘ ... that 61.82% and 56.01% of occurrences of the lemma Romanian co-occurred within a … ‘, ‘ … that it largely contributes to the negative representation … ‘, ‘ … that Romanian migration is a highly probable, large, uncontrollable, one-way phenomenon’, ‘ … that British newspapers recontextualise a perceived phenomenon …’, etc. This is also reflected in the DC/C and DC/T measures for the English L1 (0.43 and 0.77) and ESL (0.71 and 2.14) groups compared to much lower values in the EFL group (0.36 and 0.53). The analysis of the entire result sections of the dissertations of the three groups also showed much higher values of these two measures for the English L1 and ESL groups compared to the EFL group (table 6.18). This instance of form-function relationship of syntactic structures and their expected rhetorical functions based on the texts of expert writers in Lu et al. (2020) compared to these sample texts, as well as the overall quantitative findings in the dissertations, has implications for EFL writing programmes as will be discussed in chapter seven.

6.3.5. A Summary of Key Findings of 6.3

This section first demonstrated if any of the selected measures of lexical and syntactic complexity could capture statistically significant differences between the three groups of EFL, ESL, and English L1 based on different English language backgrounds and academic contexts. For the measures that showed such initial differences, further post-hoc multiple comparison tests of Tukey HSD accompanied by effect sizes and confidence intervals were administered to find the comparison sets whose mean differences mark the statistically significant difference based on the Bonferroni-corrected alpha levels in each of the six rhetorical sections.

The findings indicate that, overall, the English L1 texts followed by the ESL texts were more lexically and syntactically complex across the rhetorical sections. The differences are more noticeable in the constructs of lexical density and diversity as well as syntactic constructs of length of production units, subordination, and phrasal complexity. Overall, the English L1 and ESL groups produced larger amounts of subordination structures providing support for higher syntactic proficiency; the EFL group produced larger amounts of coordination structures that are believed to be syntactic features of the less-advanced learners. These findings are consistent with the findings of Bardovi-Harlig and Bofman (1989), Grant and Ginther (2000), Mancilla et al. (2015); Monroe (1975), Norris and Ortega (2009), and Chen, Alexopoulou, and Tsimpli, 2019) among others. The three groups did not differ significantly with regard to the production of sophisticated verbs.

With regard to the performance of specific complexity measures, the word-string-based lexical diversity measures of mattr, msttr, and mtld could consistently capture between-group differences across the rhetorical sections (apart from the literature review section), proving to be reliable indicators of lexical complexity differences of postgraduate academic writing with rhetorical functions and various text lengths. The syntactic measures of MLT (Mean Length of T-unit), C/T (Clauses per T-unit), CT/T (Complex T-units per T-unit), DC/C (Dependent Clauses per Clause), and DC/T (Dependent Clauses per T-unit) consistently captured between-group differences across various rhetorical sections claiming to be reliable indicators of syntactic complexity differences of such texts.

Other interesting patterns also emerged from the data regarding these complexity measures. For instance, the TTR-based lexical diversity measures of lv, nv, and vv2 which are more text-length dependent and have lexical tokens in their denominators, generally show larger values for the EFL group, especially in the long literature review sections; the word-string-based lexical diversity measures of mattr, msttr, and mtld, on the other hand, generally specify English L1 and ESL groups to be more lexically diverse across the rest of rhetorical sections. The largest point estimate effect sizes for the group comparisons belong to NS-EFL comparison in the abstract sections with 1.04 for msttr, followed by mattr and mtld both with 0.92 for NS-EFL comparison in the abstract sections, and msttr with 0.92 for NS-EFL as well as both mattr and mtld with 0.89 for NS-EFL in the conclusion sections.

Noticeable patterns were also observed regarding the six rhetorical sections. For instance, the rhetorical sections that are reporting and descriptive in nature (i.e., method, results, and conclusion) appeared to be more lexically dense (especially for English L1 and ESL groups) than the sections which are informational and explanatory in nature (i.e., introduction and literature review). Overall, in the rhetorical sections of abstract, introduction, results, and conclusion the English L1 then ESL groups produced more lexically dense and varied texts; in the literature review sections, however, the EFL group is shown to be more lexically diverse. The analysis of the method sections produced mixed results regarding lexical diversity. The EFL group produced more lexically-sophisticated texts (calculated as ls1) than the other two groups in the aggregated corpus as well as in the method, results, and conclusion sections; English L1 and ESL groups only produced more lexically-sophisticated texts (calculated as ls2) in the introduction sections.

Taking proficiency as a proxy to complexity, as discussed at length in 1.3 and in chapter three, the quantitative findings imply higher proficiency of the English L1 group followed by the ESL group as manifested via the larger values of the measures representative of the constructs of lexical density, diversity, and syntactic length, subordination and phrasal complexity. Even though I accompanied these results with some linguistic examples from the dissertations, future studies with detailed qualitative analyses of such texts are needed to complement the picture obtained from the quantitative analyses in this study for a better understanding of the relationship between these complexity constructs and proficiency. Having a grasp of the dissertations' lexical and syntactic complexity differences, I now focus mainly on the measure-testing process in the following sections.

6.4. Investigating the Relationship Among the Lexical and Syntactic Complexity Indices

At the second stage of the measure-testing process, I step away from group comparisons to focus on the relationship between the complexity measures. As discussed in detail in the introduction chapter as well as chapter five, by adopting a systems view of linguistic complexity, I attempt at finding the relationship between and among the selected lexical and syntactic complexity measures. This is to first examine the construct-distinctiveness of the overall lexical and syntactic categories/constructs and second to examine the extent the measures that represent each sub-construct correlate with other measures that represent other sub-constructs. These correlation tests will also help to identify sub-models in the confirmatory factor analyses as will be discussed in 6.5.1 and 6.5.3. I also attempt to compare highly-correlated measures in both datasets with the exploratory factor analysis results in 6.5.2 and 6.5.4 to examine whether such measures indeed belong to the same underlying factor/construct and whether any additional constructs could be found based on the findings of these two sections. Together with the findings of the rest of the statistics in this chapter, the relationship between these measures also helps to identify the best measure among the relevant set of measures in each sub-construct that can better discriminate lexical and syntactic complexity of the postgraduate groups' academic texts.

To find out the relationship between each set of complexity measures, three sets of Pearson correlation tests were carried out on the aggregated (entire corpus) lexical, syntactic and the lexical-syntactic datasets, using the *corrplot* package (version 0.84, Wei & Simko, 2017). The correlation matrices of coefficients as well as highly-correlated variables in each dataset are presented in tables 6.21 to 6.24 below. Table 6.25 presents the correlation coefficients for the lexical-syntactic combined data.

Weak correlation values between the three broad lexical categories of lexical density, diversity, and sophistication as indicated by table 6.21 support the assumption that they are indeed three distinct constructs. This finding is in line with the results in Lu (2012) which investigated a corpus of ESL oral narratives with many of these lexical measures. Overall, higher correlations in this study were observed between the lexical measures of the same construct, than the measures defined in different constructs. The highest positive between-construct correlations were spotted for the ld and ls1 (r = 0.4) as well as for ls2 and cvv1 (r = 0.4) and the highest negative between-construct correlations were noticed for ld and lv (r = -0.69) as well as for ld and vv2 (r = -0.5). Unlike other lexical diversity measures, the maas index was expected to produce negative values for texts with higher lexical diversity; its largest correlation value with a measure from another construct was found with cvv1 (r = -0.48).

A quick look at table 6.22 shows that most of the highly-correlated lexical measures at 0.8 and above are lexical diversity measures. It also validates the assumption that different indices that were classified into the same sub-categories of lexical diversity (e.g., word-string based, logarithm-based, TTR of word classes) have generally higher positive correlations. The highest correlation between the indices in the same sub-categories of lexical diversity was found between mattr and msttr with r = 1. This was an expected result since both indices follow very similar calculation methods; however, the effect sizes for the group differences were slightly higher for the mattr index as presented in 6.3.1. The next highest correlation was observed between the two lexical sophistication indices of vs2 and cvs1 with r = 0.98. None of these two indices, however, captured between-group differences in the analyses of variance in the previous section which could be an indication that the three groups of English L1, ESL, and EFL did not produce significantly different number of sophisticated verbs as described in 6.3.1.

Tables 6.23 and 6.24 reveal very interesting findings about the correlations between syntactic measures: every syntactic index included in this study has at least one high correlation with another index at 0.8 or above. However, most of these high correlations are spotted between the measures in the same constructs (e.g., length of structures, subordination, coordination, phrasal sophistication) and generally lower correlations were found between the indices from different categories/constructs. This finding supports the assumption that such categories are indeed distinct constructs as defined by Lu (2010) as well as Lu and Ai (2015). The highest between-construct correlation is shown for MLT and CN/T at r = 0.93 and the lowest negative between-construct correlation is found for CP/C and C/T with r = -0.2 as well as between CP/C and CT/T with r = -0.22. In the exploratory factor analysis section in this chapter, I will further explore whether the high correlation between MLT and CN/T also results in their similar factor loadings on one factor/construct.

Table 6.21. Correlation matrix for 22 lexical complexity indices in the entire corpus

	maas	ld	ls1	ndwesz	mtld	mattr	msttr	ndwerz	hdd	vocd	logttr	rttr	uber	ls2	cvv1	vs2	cvs1	nv	lv	vv2	vv1	adjv
maas	1																					
ld	-0.05	1																				
ls1	0.07	0.44	1																			
ndwesz	-0.7	-0.07	-0.11	1																		
mtld	-0.77	0.0	-0.11	**0.85**	1																	
mattr	-0.79	-0.0	-0.14	**0.9**	**0.96**	1																
msttr	-0.79	-0.0	-0.14	**0.89**	**0.96**	**1**	1															
ndwerz	-0.63	0.01	0.01	0.69	0.67	0.71	0.71	1														
hdd	-0.7	-0.08	-0.08	0.76	**0.82**	**0.86**	**0.86**	**0.83**	1													
vocd	-0.63	0.0	-0.01	0.69	0.79	0.77	0.77	0.77	**0.89**	1												
logttr	**-0.83**	0.04	0.0	0.54	0.59	0.58	0.58	0.43	0.47	0.48	1											
rttr	**-0.89**	0.01	-0.002	0.59	0.67	0.66	0.66	0.58	0.62	0.6	0.63	1										
uber	**-0.96**	0.04	0.0	0.64	0.73	0.71	0.7	0.59	0.62	0.63	**0.86**	**0.93**	1									
ls2	-0.36	0.2	0.33	0.14	0.26	0.18	0.18	0.28	0.2	0.32	0.13	0.64	0.47	1								
cvv1	-0.48	-0.21	-0.18	0.29	0.33	0.35	0.35	0.38	0.33	0.3	0.24	0.57	0.48	0.42	1							
vs2	-0.4	-0.08	0.03	0.21	0.28	0.27	0.27	0.35	0.28	0.28	0.18	0.56	0.44	0.55	**0.84**	1						
cvs1	-0.42	-0.08	0.05	0.23	0.29	0.28	0.28	0.37	0.3	0.29	0.19	0.57	0.45	0.57	**0.86**	**0.98**	1					
nv	-0.33	-0.4	0.04	0.18	0.19	0.18	0.19	0.08	0.15	0.14	0.59	0.18	0.38	-0.11	-0.02	-0.05	-0.04	1				
lv	-0.37	-0.69	-0.33	0.28	0.25	0.27	0.27	0.17	0.22	0.18	0.55	0.23	0.39	-0.12	0.37	0.2	0.2	0.73	1			
vv2	-0.28	-0.56	-0.07	0.09	0.1	0.11	0.11	0.09	0.08	0.0	0.4	0.21	0.32	0.03	0.38	0.33	0.36	0.63	0.76	1		
vv1	-0.16	-0.33	-0.37	0.16	0.15	0.15	0.15	0.13	0.13	0.16	0.24	0.05	0.15	-0.13	0.63	0.35	0.36	0.11	0.55	0.22	1	
adjv	-0.3	-0.22	-0.22	0.26	0.22	0.23	0.22	0.22	0.21	0.26	0.37	0.2	0.31	-0.1	0.23	0.04	0.03	0.29	0.43	0.17	0.41	1

– Unlike other lexical diversity indices, the maas index shows negative values for higher lexical diversity; therefore, large negative correlation values of the maas index with any other measure denote stronger correlation.

Table 6.22. Highly-correlated lexical measures in the entire corpus

Correlation coefficients	**.8 < r < .9**	**R => .9**
Pairs of correlated measures	maas vs. logttr -0.83 maas vs. rttr -0.89 logttr vs. uber 0.86 ndwesz vs. mtld 0.85 ndwesz vs. msttr 0.89 mtld vs hdd 0.82 mattr vs. hdd 0.86 msttr vs. hdd 0.86 ndwerz vs. hdd 0.83 hdd vs. vocd 0.89 cvv1 vs. vs2 0.84 cvv1 vs. cvs1 0.86	maas vs. uber -0.96 rttr vs. uber 0.93 ndwesz vs. mattr 0.90 mtld vs. mattr 0.96 mtld vs. msttr 0.96 mattr vs. msttr 1 vs2 vs. cvs1 0.98

Table 6.23. Correlation matrix for 11 syntactic complexity indices in the entire corpus

	VP/T	**C/T**	**DC/T**	**CT/T**	**DC/C**	**CP/C**	**CP/T**	**MLT**	**CN/T**	**MLC**	**CN/C**
VP/T	1										
C/T	**0.86**	1									
DC/T	**0.89**	**0.98**	1								
CT/T	**0.85**	**0.93**	**0.95**	1							
DC/C	**0.87**	**0.91**	**0.97**	**0.95**	1						
CP/C	0.01	-0.2	-0.18	-0.22	-0.17	1					
CP/T	0.4	0.25	0.26	0.2	0.24	**0.89**	1				
MLT	0.78	0.68	0.7	0.64	0.67	0.33	0.63	1			
CN/T	0.61	0.55	0.56	0.51	0.54	0.32	0.57	**0.93**	1		
MLC	0.16	-0.1	-0.06	-0.08	-0.03	0.66	0.61	0.65	0.69	1	
CN/C	0.07	-0.08	-0.05	-0.08	-0.04	0.54	0.5	0.6	0.78	**0.9**	1

Table 6.24. Highly-correlated syntactic measures in the entire corpus

Correlation coefficients	**.8 < r < .9**	**r => .9**
Pairs of correlated measures	VP/T vs. C/T 0.86 VP/T vs. DC/T 0.89 VP/T vs. CT/T 0.85 VP/T vs. DC/C 0.87 CP/C vs. CP/T 0.89	C/T vs. DC/T 0.98 C/T vs. CT/T 0.93 C/T vs. DC/C 0.91 DC/T vs. CT/T 0.95 DC/T vs. DC/C 0.97 CT/T vs. DC/C 0.95 MLT vs. CN/T 0.93 MLC vs. CN/C 0.90

Table 6.25. Correlation matrix for 22 lexical and 1ll syntactic complexity indices in the entire corpus

	maas	**ld**	**ls1**	**ndwesz**	**mtld**	**mattr**	**msttr**	**ndwerz**	**hdd**	**vocd**	**logttr**	**rttr**	**uber**	**ls2**	**cvv1**	**vs2**	**cvs1**	**nv**	**lv**	**vv2**	**vv1**	**adjv**
MLC	-0.20	0.02	0.22	0.16	0.16	0.15	0.15	0.13	0.08	0.03	0.16	0.27	0.24	**0.43**	0.15	0.2	0.21	0.18	0.11	0.21	-0.08	-0.06
MLT	-0.05	-0.02	-0.008	0.18	0.19	0.21	0.21	0.04	0.05	0.002	-0.04	0.06	0.02	0.18	0.09	0.08	0.08	0.01	0.03	0.09	-0.06	-0.08
C/T	0.12	-0.04	-0.22	0.09	0.09	0.14	0.14	-0.05	0.0004	-0.02	-0.23	-0.17	-0.21	-0.17	-0.02	-0.08	-0.09	-.18	-.08	-0.09	-0.006	-0.03
CT/T	0.11	0.02	-0.17	0.13	0.15	0.19	0.19	-0.01	0.02	-0.01	-0.25	-0.17	-0.21	-0.16	-0.01	-0.07	-0.07	-.23	-.14	-0.15	-0.01	-0.08
DC/C	0.10	0.01	-0.13	0.12	0.12	0.17	0.18	-0.006	0.03	-0.01	-0.23	-0.16	-0.20	-0.16	-0.03	-0.09	-0.09	-.18	-.12	-0.12	-0.05	-0.05
DC/T	0.13	-0.01	-0.18	0.09	0.09	0.14	0.14	-0.03	0.01	-0.03	-0.25	-0.19	-0.22	-0.17	-0.03	-0.10	-0.11	-.18	-0.11	-0.12	-0.03	-0.05
CP/C	-0.22	-0.08	0.18	0.21	0.19	0.18	0.18	0.23	0.22	0.16	0.15	**0.34**	0.28	**0.41**	0.21	0.22	0.24	0.12	0.11	0.18	-.0005	0.06
CP/T	-0.16	-0.12	0.05	0.25	0.24	0.25	0.25	0.20	0.22	0.15	0.04	0.26	0.18	**0.31**	0.19	0.17	0.19	0.05	0.08	0.15	-0.001	0.05
CN/C	-0.13	0.04	0.27	0.04	0.08	0.06	0.06	0.05	0.001	-0.03	0.16	0.21	0.20	**0.47**	0.11	0.18	0.19	0.24	0.12	0.22	-0.10	-0.12
CN/T	-0.04	0.001	0.08	0.09	0.13	0.14	0.14	0.005	-0.001	-0.04	-0.003	0.07	0.04	0.27	0.08	0.10	0.10	0.10	0.06	0.13	-0.08	-0.13
VP/T	-0.01	-0.13	-0.17	0.27	0.24	**0.30**	0.29	0.10	0.18	0.08	-0.12	-0.04	-0.07	-0.16	0.001	-0.05	-0.06	-.04	.004	-0.02	-0.006	0.001

– The highest correlation between lexical and syntactic measures are ls2 vs. CN/C (r = 0.47), ls2 vs. MLC (r = 0.43), and ls2 vs. CP/C (r = 0.41).

The highest correlation between the measures in the same construct/category is between C/T (clauses per T-unit) and DC/T (dependent clauses per T-unit) at r = 0.98. This was an expected result since both indices calculate similar ratios; however, DC/T measure captured the between-group differences with larger effects as indicated in the previous section. Another interesting finding was the lower-than-expected correlation between the two indices calculating the length of production, namely MLT and MLC (r = 0.6), while other indices belonging to the same construct generally showed correlations above 0.8. As will be discussed in the results of exploratory factor analysis in section 6.5.4, MLC is also loaded alone on a separate factor, denoting a distinct syntactic construct, while MLT is loaded alongside other subordination indices.

Regarding the combined data, overall, trivial correlations were obtained in the lexical-syntactic combined datasets. This shows that various lexical and syntactic indices do not have any (meaningful) relationships, suggesting that they indeed tap different overall constructs/dimensions of proficiency. However, a few interesting and noteworthy patterns were observed. The highest coefficients between lexical and syntactic measures are found for the correlations between ls2 (lexical sophistication type II) and CN/C (complex nominals per clause), MLC (mean length of clauses), and CP/C (coordinate phrases per clause) which have the number of clauses as their denominators. Even though these coefficients are not statistically significant (e.g., they are not above 0.7), they show an interesting pattern regarding the co-occurrence of sophisticated lexical items and these specific syntactic structures, specifically complex nominals. The presence of abundant nominal structures such as complex nominals is believed to be an indicator of higher levels of academic writing. In this study, lexical sophistication indices, including ls2, are more field-specific, i.e., are filtered through frequently-used words in linguistics-related disciplines. This finding suggests that complex nominals contain more of such sophisticated lexical words than other syntactic structures. A few examples of these lexically-sophisticated items within complex nominal structures in the excerpts in 6.3.2 and 6.3.4, are ‘that British newspapers **recontextualise** ...’ (in a nominal clause), ‘**manuals** of instruction’ (in N + PP), ‘reactions to **cooperative** learning’ (in N + PP), and ‘**elicitation** studies on Arab learners which conclude that ...’ (in N + relative clause), etc.

6.4.1. A Summary of Key Findings of 6.4

This section examined the relationship between and among lexical and syntactic complexity measures. Weak correlations were found between the lexical measures representing different constructs of lexical density, diversity and sophistication; the same pattern is observed for the syntactic constructs of subordination, coordination and phrasal sophistication. These findings indicate that the mentioned categories are indeed different constructs. Higher correlations were noticed between the lexical measures specified in the same constructs and then between the same sub-categories of the same constructs. These findings are further supported by the constructs that were detected in the exploratory factor analyses (see section 6.5 below), suggesting that the indices defined in each category indeed measure the same constructs.

Most of the highly-correlated lexical measures at r >= 0.8 are lexical diversity measures. Among these measures, logarithm-based and word-string-based indices of lexical diversity have much higher correlations between and among each other than with the lexical diversity of TTR of word classes. Among the syntactic indices included in this study, each index has at least one high correlation with another index at 0.8 or above. We also notice a lower-than-expected correlation between the two indices measuring the length of production, namely MLT and MLC (r = 0.6). The results of factor analysis further confirms this point, where the MLC index is loaded alone on as a separate factor.

An interesting finding, though not statistically significant, is the correlation and the co-occurrence of ls2 as lexical sophistication measure with the three syntactic measures of CN/C, MLC, and CP/C with the number of clauses as their denominators. This result suggests that in this study’s discipline-specific postgraduate corpus, complex nominals, for instance, contain more lexically-sophisticated words, followed by the syntactic structures of clauses and coordinate phrases.

In the next section, the results of these correlation tests will be used to specify models and sub-models mainly to avoid model convergence issues that can be caused by high correlations among various measures that belong to different constructs as will be discussed in detail.

6.5. Structural Factor Analysis: Detecting the Structure of Lexical and Syntactic Datasets

An important next step in the measure-testing process and after finding the relationship between lexical and syntactic complexity measures was to find the overall and specific structure of the two datasets to compare the proposed structure of these measures (e.g., in Lu, 2012 and Lu & Ai, 2015) with this study’s corpus of master’s dissertations. The correlations between and among lexical and syntactic complexity measures and constructs so far helped to obtain a picture of the boundaries between the constructs in each dataset and to find out the amount of correlation between the indices as specified in 6.4. This is useful to assess which specific lexical and syntactic measures are correlated with each other. However, it does not constitute a formal test of the idea that the different variables actually constitute one underlying construct, for which a Structural Factor Analysis is needed. A series of factor analyses were, therefore, used to verify the structure of such constructs and how well the variables represent each construct based on the proposed structures and to further explore the datasets for any additional construct/latent factor that can be revealed, or any misplacement of measures based on the postgraduate academic corpus at hand. Since these structures or classifications are not formally proposed as ‘models’, in this study I use the word ‘model’ as a proxy for these structures/classifications in the confirmatory factor analysis phase. A series of factor analyses (FA) was conducted to examine these proposed lexical and syntactic structures of constructs (Confirmatory Factor Analysis, CFA) and to find the clusters of similar/homogeneous lexical and syntactic measures and to investigate whether there are latent factors among different measures, e.g., statistical evidence for a theoretical construct (Exploratory Factor Analysis, EFA). Factor analyses also examine the validity of constructs and measures based on the proposed structures in the literature and their conceptual and theoretical assumptions.

Prior to this stage, a number of pre-requisite statistics were carried out to assess the assumptions of factor analysis. The first is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (Kaiser, 1974) which shows the proportion of variance in the complexity measures that could be caused by underlying factors. Higher values of MSA (Measure of Sampling Adequacy), ideally closer to 1 and more than 0.50, suggest that factor analysis is useful/justified. This statistic was computed using the *KMO* function in the *psych* package (version 1.8.12, Revelle, 2018).

Table 6.26. KMO test on the aggregated lexical dataset

Lexical Measures	**ld**	**ls1**	**ls2**	**vs2**	**cvs1**	**ndwerz**	**ndwesz**	**maas**	**logttr**	**uber**	**rttr**
MSA	0.59	0.61	0.83	0.77	0.80	0.94	0.93	0.81	0.80	0.83	0.84
Lexical Measures	**lv**	**vv1**	**cvv1**	**vv2**	**nv**	**adjv**	**mattr**	**msttr**	**mtld**	**Vocd**	**hdd**
MSA	0.55	0.77	0.65	0.86	0.88	0.91	0.86	0.88	0.86	0.92	0.88

– The overall MSA (Measure of Sampling Adequacy) for the lexical dataset = 0.83

The results of the overall KMO tests on both lexical dataset (table 6.26, MSA = 0.83) and syntactic dataset (table 6.27, MSA = 0.66) show the amount of variance that might be caused by underlying factors is beyond the recommended threshold e.g., the suggested 0.50 (see Yoon, 2017). This indicates that conducting factor analysis is justified.

Table 6.27. KMO test on the aggregated syntactic dataset

Syntactic Measures	**MLC**	**MLT**	**C/T**	**CT/T**	**DC/C**	**DC/T**	**CP/C**	**CP/T**	**CN/C**	**CN/T**	**VP/T**
MSA	0.44	0.58	0.68	0.89	0.75	0.76	0.60	0.67	0.46	0.58	0.97

– The overall MSA (Measure of Sampling Adequacy) for the syntactic dataset = 0.66

The next pre-requisite is Bartlett's Test of Sphericity (explained in Tobias & Carlson, 2010) which tests if the correlation matrix deviates from an identity matrix (i.e., a matrix where all diagonal elements are 1 and all other elements are 0). Higher values of the chi-square statistic with low values of the associated p-value (lower than 0.05) suggest that there are some strong correlations between some variables and therefore, the data is suitable for factor analysis and structure detection (Tobias & Carlson, 2010; Yoon, 2017). This test was done using the *psych* package. The results of Bartlett's test on both datasets (table 6.28) also show high values of chi-square tests with very small p-values which suggest that factor analysis and structure detection is possible in both datasets.

Table 6.28. Bartlett's test on the aggregated lexical and syntactic datasets

Bartlett's Test	**Chi-Square**	**P-value**	**Degrees of Freedom (df)**
Lexical Dataset	9257.37	0	300
Syntactic Dataset	6307.62	0	5

McArdle (2011) recommends an 'ethical' practice for factor analysis named Structural Factor Analysis (SFA) which is partly based on the proposition of Cattell (1966; cited in McArdle, 2011) of a continuum beginning with confirmatory factor analysis (CFA) and ending with exploratory factor analysis (EFA). By adopting this 'confirm first, explore second' procedure, he insists, one can "avoid the artificial semantic differences of confirmation and exploration" as exploited by many researchers in unethical ways, and improve the model without manipulating the statistics to obtain the desired outcome (McArdle, 2011, pp. 315-333). This practice starts with a CFA based on existing theories and/or a priori knowledge of variables, their constructs and their overall and specific structures that are proposed in previous scholarship. In the case of this study, CFA confirms/assesses the structure of the constructs and their representative measures. We then proceed with an exploratory factor analysis to examine if any additional construct could be detected (e.g., any additional construct that has not been accounted in the previous classifications in the literature).

For the CFA tests, McArdle (2011) suggests a Maximum Likelihood Estimation (MLE) using Structural Equation Modelling (SEM) computer algorithms. The model's fit is then compared to the existing models/theories/constructs and if significant differences are found (e.g., the misfit of a model as indicated by goodness-of-fit measures), the model and/or theories could be re-evaluated. This step is then followed by a series of exploratory factor analysis tests to explore the data further and to find any latent variable that we did not take into account in the CFA step (e.g., the factors or constructs that were not proposed in the literature). The researcher records the analyses' indices of model fit/indices of goodness of fit (e.g., RMSEA, Chi-square and degrees of freedom, RMSR, and Tucker-Lewis Index of reliability) to arrive at a model with the best fit. This final EFA model is consequently compared with the initial CFA model and the model could be improved further either by a subsequent final CFA model and/or by the judgement of the researcher about the nature of the variables, existing theories, and a comparison of the analyses.

The following sections will investigate lexical and syntactic datasets using the CFA and EFA tests with detailed information on the methods of conducting these tests and the evaluation scheme. To avoid extra technical details in the following sections and to aid the reading flow, a detailed explanation of the evaluation scheme using the model fit indices will be presented in Appendix C1. Since the tests in section 6.5 mainly cater for the measure-testing process, most of the following discussions will be based on the behaviours of individual measures rather than a purely linguistic interpretation.

In this study, the confirmatory factor analysis was done using the *lavaan* package (version 0.5-18, Rosseel, 2012). This method uses the *sem* function with the maximum likelihood estimation (MLE) as suggested by McArdle (2011). A robust estimator was also used to compensate for any deviation from normality in the values of the measures. For each dataset, first, an analysis was carried out with the full dataset including all variables. If the results showed a misfit, two parallel models were conducted whereby each measure in a pair of highly-correlated variables (e.g., r >= 0.9) is included in one model but not in the other.

This is to ascertain if the misfit of the initial CFA test is a genuine misfit or due to high correlations among measures of different constructs (e.g., by noticing the error/warning messages in each test). The assumption is that if one variable is shown to belong to a specific factor, another highly-correlated variable with it would consequently belong to the same factor/underlying construct. I will investigate whether any of the two parallel models produce significantly better fit indices (i.e., dropping highly-correlated variables could be an effective method to avoid multicollinearity issues in the initial model) or whether both models result in similar overall fit indices (i.e., the initial model's misfit is a genuine misfit).

Since the measures in this study have different scales (i.e., different numeric ranges/ different metric), I standardised the measures using the z-score method of standardisation (i.e., by subtracting the mean and dividing by the standard deviation of any measure).

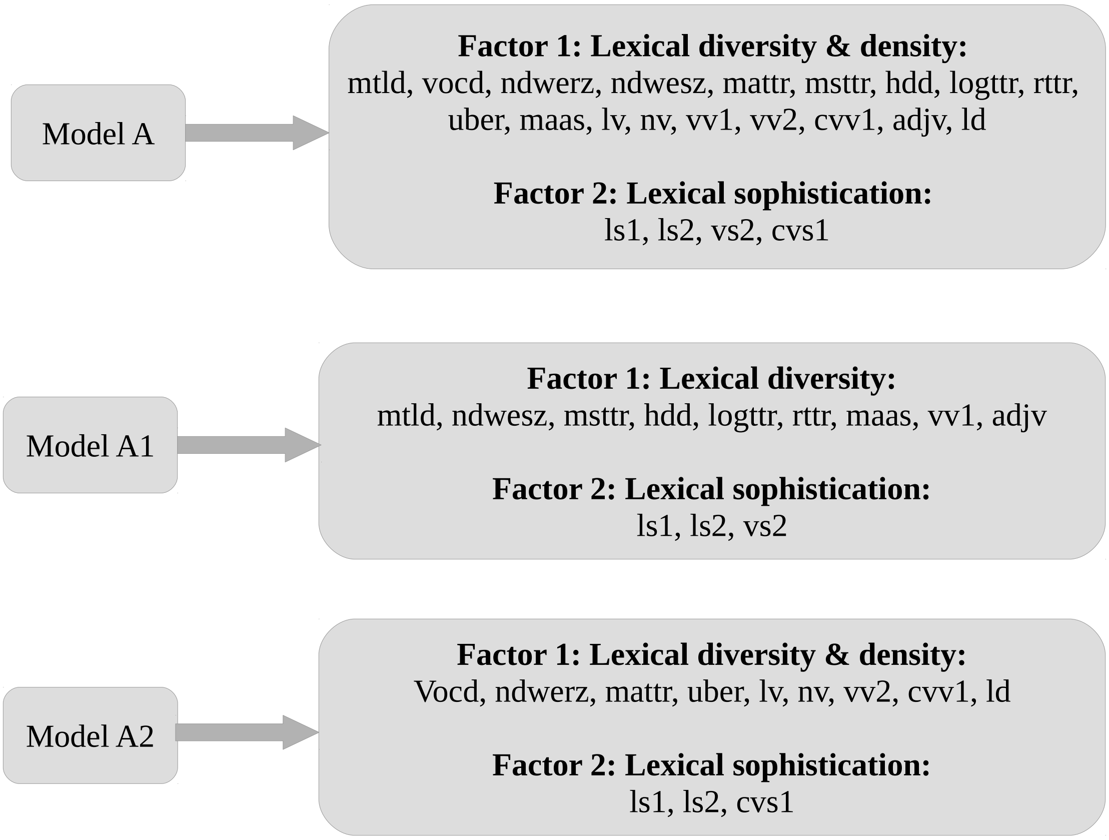
In the next step, a series of exploratory factor analyses were conducted using the 'maximum likelihood' extraction/factoring method. Since I expected some degrees of correlation to be found between variables, I used an 'oblique' rotation, specifically the 'direct oblimin' rotation type as implemented in the *psych* package. The oblique rotation is recommended when it is theoretically plausible that the factors are correlated with each other. I initially determined the number of factors to start with, based on the *fa.parallel* function in the *psych* package together with the scree plot and its suggested eigenvalues greater than 1 (see for example the discussions in Costello & Osborne, 2005). The goodness of fit indices of chi-square, degrees of freedom (df), Tucker-Lewis Index of reliability (TLI), RMSR, and RMSEA were recorded at each step as suggested by McArdle (2011), and Kline (2005) among others. The best-fitting model was compared to the CFA results to interpret the model and to find new aspects of the constructs and the measures. Detailed information about these goodness-of-fit indices and the ways to interpret the results based on them are presented in Appendix C1.

6.5.1. Confirmatory Factor Analysis of the Lexical Data

The literature on lexical complexity and/or richness offers two types of classification of lexical indices, one a general model where all types of commonly-understood and reported lexical diversity indices are assumed to belong to one overall construct (e.g., the discussions in chapter two, section 2.2.2), and the other one with a more fine-grained classification (see for instance Lu, 2012 and the classification in this thesis in 5.3.1.1) where different lexical diversity measures are allocated different sub-constructs/sub-categories (e.g., logarithm-based indices, indices based on word strings/segments, lexical variation based on the type-token ratio of word classes). Since in this research I investigated an extended set of lexical diversity measures compared to Lu (2012) the rest of the measures that were not investigated in Lu's study are allocated a sub-category based on the previous works that were discussed in chapters two and five.

Based on these specifications and the explanations in the previous section, two types of lexical models are tested based on the mentioned criteria. The general Model A examines the underlying constructs, i.e., what is generally understood as lexical diversity versus lexical sophistication and density in the literature, and Model B which tests the fine-grained (i.e., sub-categories) version of Model A. The latter model is included to examine whether the sub-categories of lexical diversity indeed belong to the same construct based on this study's specialised academic writing corpus. For each model type, two parallel models are specified with the same number of factors as their base models but with fewer indicators compared to their base models. The specifications of Models A and B with their sub-models are as follows:

Diagram 6.1. General lexical model A and its two sub-models A1 and A2 with omitted highest-correlated variables



Model A

Factor 1: Lexical diversity & density:
mtld, vocd, ndwerz, ndwesz, mattr, msttr, hdd, logttr, rttr, uber, maas, lv, nv, vv1, vv2, cvv1, adjv, ld

Factor 2: Lexical sophistication:
ls1, ls2, vs2, cvs1

Model A1

Factor 1: Lexical diversity:
mtld, ndwesz, msttr, hdd, logttr, rttr, maas, vv1, adjv

Factor 2: Lexical sophistication:
ls1, ls2, vs2

Model A2

Factor 1: Lexical diversity & density:
Vocd, ndwerz, mattr, uber, lv, nv, vv2, cvv1, ld

Factor 2: Lexical sophistication:
ls1, ls2, cvs1

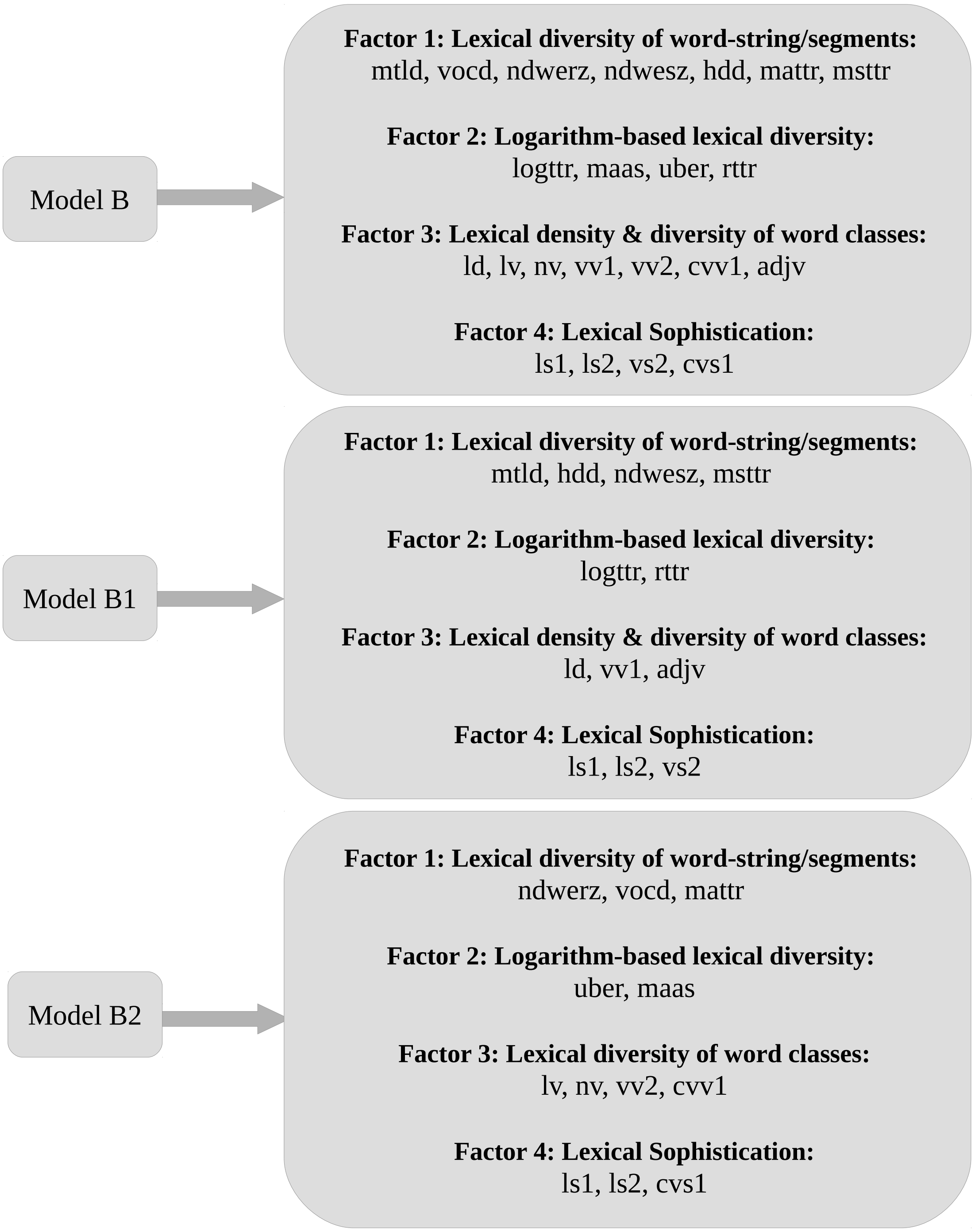
--The arrows only denote assignment

However the above model needed some modification due to the test's logistic reasons: since CFA tests perform better with at least two indicators per latent variable (i.e., single indicator latent variables/factors are not recommended; see for instance the discussions in Kline, 2005), I moved the ld or lexical density measure to the next factor, expecting that I would receive a negative factor loading for this measure. This step is applied to the initial tests for both lexical models types A and B on the aggregated lexical dataset as shown in the following paragraphs. Therefore, instead of a three-factor model, I test a two-factor model as shown above. Similarly, because the maas index produces larger negative values for higher lexical diversity (as opposed to other lexical diversity indices where higher lexical diversity produces larger positive values), it is expected that this index also shows a negative factor loading on both model types.

The tests for the two parallel models (A1 and A2) with omitted highest-correlated variables, were then executed with the following model specifications; each with two factors, factor 1 with nine indicators, and factor 2 with three. That is, the number of variables in each sub-model is kept exactly the same to prevent unwanted variation in the results and for a better comparison of the two sub-models. This necessitated using either of the verb-based lexical sophistication measures (vs2 and cvs1) in each sub-model because of their very high correlation (r = 0.98 as shown in table 6.20). This was not the case for the other two sophistication measures of ls1 and ls2 since they only correlate at r = 0.33 (see table 6.19 for all correlation coefficient values that were taken into account when specifying the lexical sub-models).

As specified in the model in diagram 6.2, in Model B also, ld or lexical density is moved to the next factor, that is the lexical diversity of word classes. This also means that a negative factor loading for the ld measure is expected. Therefore, instead of testing a five-factor model, I test a four-factor model. For this model root TTR index (labelled as 'rttr') is also placed in the second factor with the logarithm-based measures because of its positive correlation with these measures as specified in table 6.21. The tests for the two parallel models with omitted highest-correlated variables were conducted with the following model specification: each with four factors and 12 indicators/measures. Because of the overall high correlation between many lexical diversity indices, the sub-models prioritise omitting the highest-correlated measures first. This means, inevitably, some high correlations will remain between measures in various factors in these two sub-models. The classification of these lexical diversity measures are both theoretically driven (e.g., based on the mathematical formulas as discussed in chapter two, section 2.2.2) and based on the existing linguistic classification (e.g., in Lu, 2012).

Diagram 6.2. Fine-grained lexical model B and its two sub-models B1 and B2 with omitted highest-correlated variables



Model B

Factor 1: Lexical diversity of word-string/segments:
mtld, vocd, ndwerz, ndwesz, hdd, mattr, msttr

Factor 2: Logarithm-based lexical diversity:
logttr, maas, uber, rttr

Factor 3: Lexical density & diversity of word classes:
ld, lv, nv, vv1, vv2, cvv1, adjv

Factor 4: Lexical Sophistication:
ls1, ls2, vs2, cvs1

Model B1

Factor 1: Lexical diversity of word-string/segments:
mtld, hdd, ndwesz, msttr

Factor 2: Logarithm-based lexical diversity:
logttr, rttr

Factor 3: Lexical density & diversity of word classes:
ld, vv1, adjv

Factor 4: Lexical Sophistication:
ls1, ls2, vs2

Model B2

Factor 1: Lexical diversity of word-string/segments:
ndwerz, vocd, mattr

Factor 2: Logarithm-based lexical diversity:
uber, maas

Factor 3: Lexical diversity of word classes:
lv, nv, vv2, cvv1

Factor 4: Lexical Sophistication:
ls1, ls2, cvs1

--The arrows only denote assignment

The results of confirmatory factor analysis on the two main lexical models, each with two sub-models are presented in Table 6.29. All models use the lexical dataset with 210 observations and all converged normally. All models were tested using the same code using the *lavaan* package and the robust statistic version of maximum likelihood, called MLM in *lavaan* as the estimator, which produces robust standard errors, and a Satorra-Bentler scaled test statistic which controls any deviations from normality in the indices' values.

Table 6.29. The comparison of lexical models' fit indices in confirmatory factor analysis

Lexical Models	**Number of Free Parameters**	**Model Chi-Square (p-value)**	**df**	**RMSEA [CIs]**	**SRMR**	**Tucker-lewis Index/ TLI**
Model A	45	3398.553 (0.000)	208	.29 [.28, .30]	.19	.45
Model A 1	25	794.552 (0.000)	53	.27 [.26, .29]	.15	.57
Model A 2	25	1113.593 (0.000)	53	.33 [.31, .34]	.22	.21
Model B	50	1989.073 (0.000)	203	.23 [.22, .24]	.16	.65
Model B1	30	280.347 (0.000)	48	.16 [.15, .18]	.13	.79
Model B2	30	638.722 (0.000)	48	.26 [.25, .28]	.17	.57

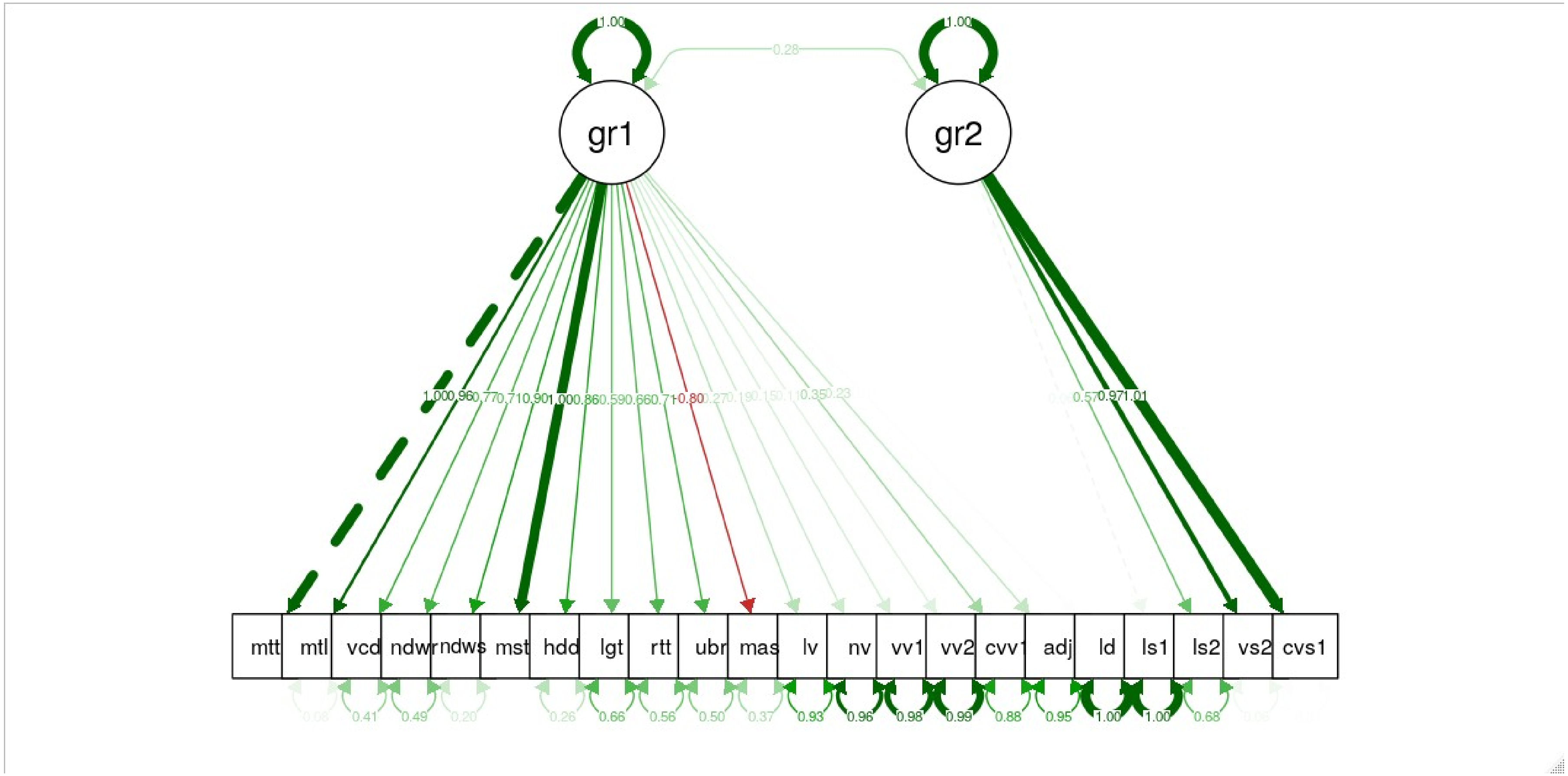
The cut-off criteria for the interpretation of goodness-of-fit indices in this table are based on Kline (2005), and Hu and Bentler (1999). To avoid extra technical notes, I only provide brief explanations of these model fit indices in this section (see Appendix C1 for more detail). The second column of table 6.29 specifies the number of free parameters which include the variances (each variance for each of the lexical indices), regression coefficients, and covariances among variables. The third column shows the results of the chi-square tests and their associated p-values. In the case of CFA models, this statistic is also sometimes called a 'badness-of-fit' index because smaller values are more desirable, e.g., denote better model fit. Chi-square statistics are sensitive to high correlations between variables/measures specified in each factor. The fourth column shows the degrees of the freedom of each model calculated internally by the *sem* function. The fifth column shows the RMSEA (Root Mean Square Error of Approximation) fit index with its confidence intervals. It measures if the model can closely reproduce the data patterns; values smaller than .08 indicate good model fit (ideally less than .06; zero indicates the perfect fit). The next column demonstrates the SRMR fit index which is the Standardised Root Mean Square Residual. This index transforms the predicted and sample covariance matrices to correlation matrices, so the values are the differences between the observed and predicted correlations. Values smaller than .08 is considered a good fit. In an ideal model, the residuals are close to zero. The last column of this table shows the Tucker-Lewis Index (TLI) of reliability. Larger values closer to .9 represent a good model. The following paragraphs discuss the results mainly from a measure-testing point of view.

A glance at table 6.29 shows that none of the three lexical models (Model A and its two sub-models of A1 and A2) produced acceptable values of fit indices as recommended by the mentioned scholars. However, among the three models, model A1 resulted in slightly better fit indices of Chi-square, RMSEA, SRMR, and TLI. This result shows that for a two-factor general lexical model, the combination of mtld, ndwesz, msttr, hdd, logttr, rttr, maas, vv1, and adjv indices as indicators of lexical diversity results in a better model than the combination of vocd, ndwerz, mattr, uber, lv, nv, vv2, cvv1, ld. Graphs 6.5, 6.6, and 6.7 portray graphical representations of the structures of models A, A1 and A2 respectively along with the factor loadings. The darker green arrows show stronger indicators for each group/factor. The standardised factor loadings of the measures (long arrows in the diagrams) are also indicated in the caption of each graph for easier reference. The red lines show negative loadings. As expected the ld and maas indices are indicated by red lines/negative loadings since the ld measure is deliberately dislocated from its own factor to avoid a single-indicator factor, and the maas index shows negative values for higher lexical diversity. The results also suggest that in the general lexical model (model A) which does not differentiate between various sub-constructs of lexical diversity, the indices in the first factor (labelled as ‘gr1’ in the diagram) do not equally represent the overall construct of lexical diversity and with the same/similar strength. This finding, however, needs to be further investigated with different academic writing corpora and possibly with larger sample size.

The dashed lines in these three diagrams represent the fixed indicator (fixed to 1), i.e., the first indicator specified in each factor for each model. This is because in such models, a constraint is applied to one variable (usually the first indicator/variable) which acts as a reference point for the model to estimate the rest of variables in terms of their relation to the latent factor (e.g., gr1 or factor 1 in these models).

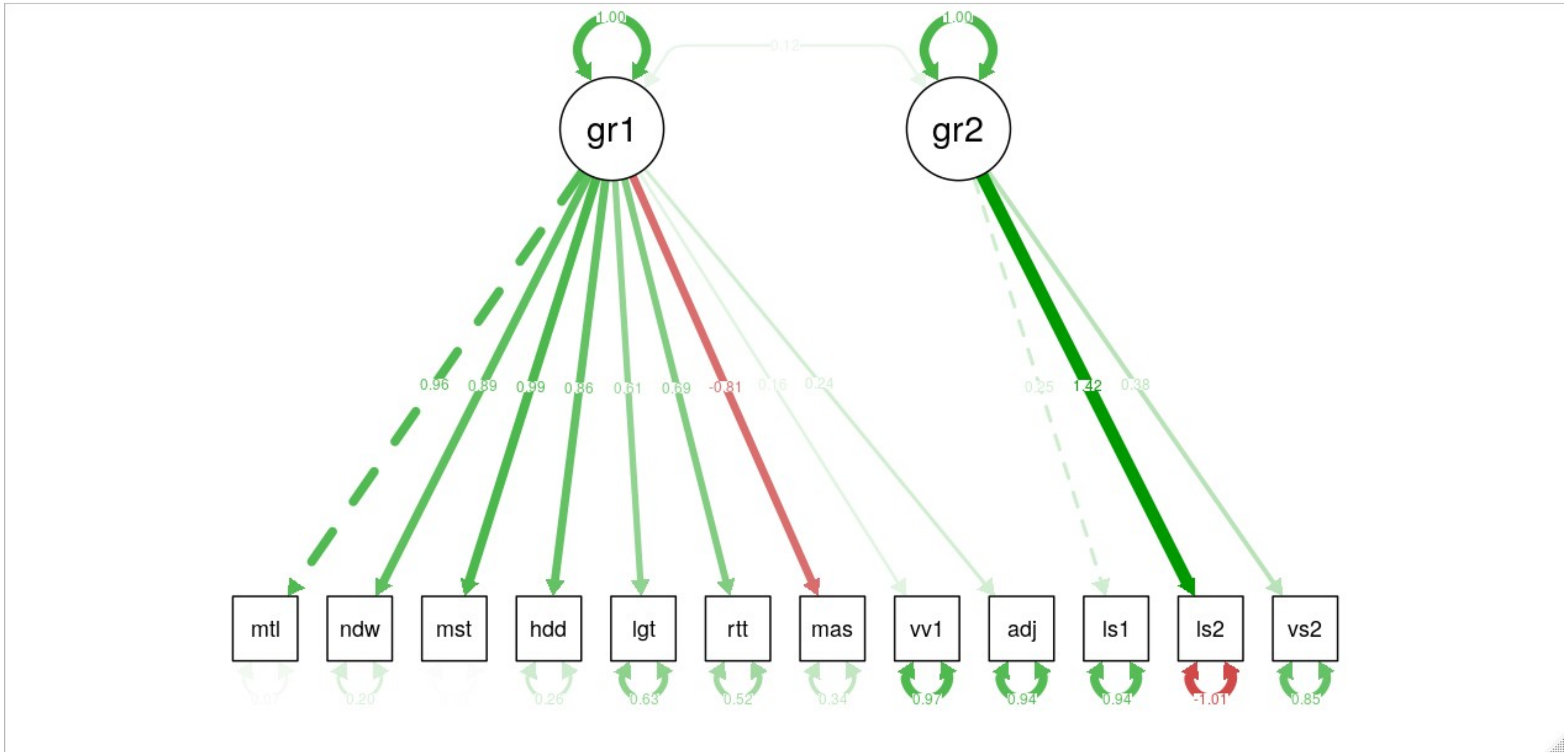
As is demonstrated in the diagram of model A, the mtld and msttr indices better represent the lexical diversity construct (factor 1 in the model specification) and the lexical variation indices based on the type-token ratio of word classes (the last six indicators in factor 1/ gr1) are the weakest representatives of this construct, in relation to all variables specified in factor 1.

Graph 6.5. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model A



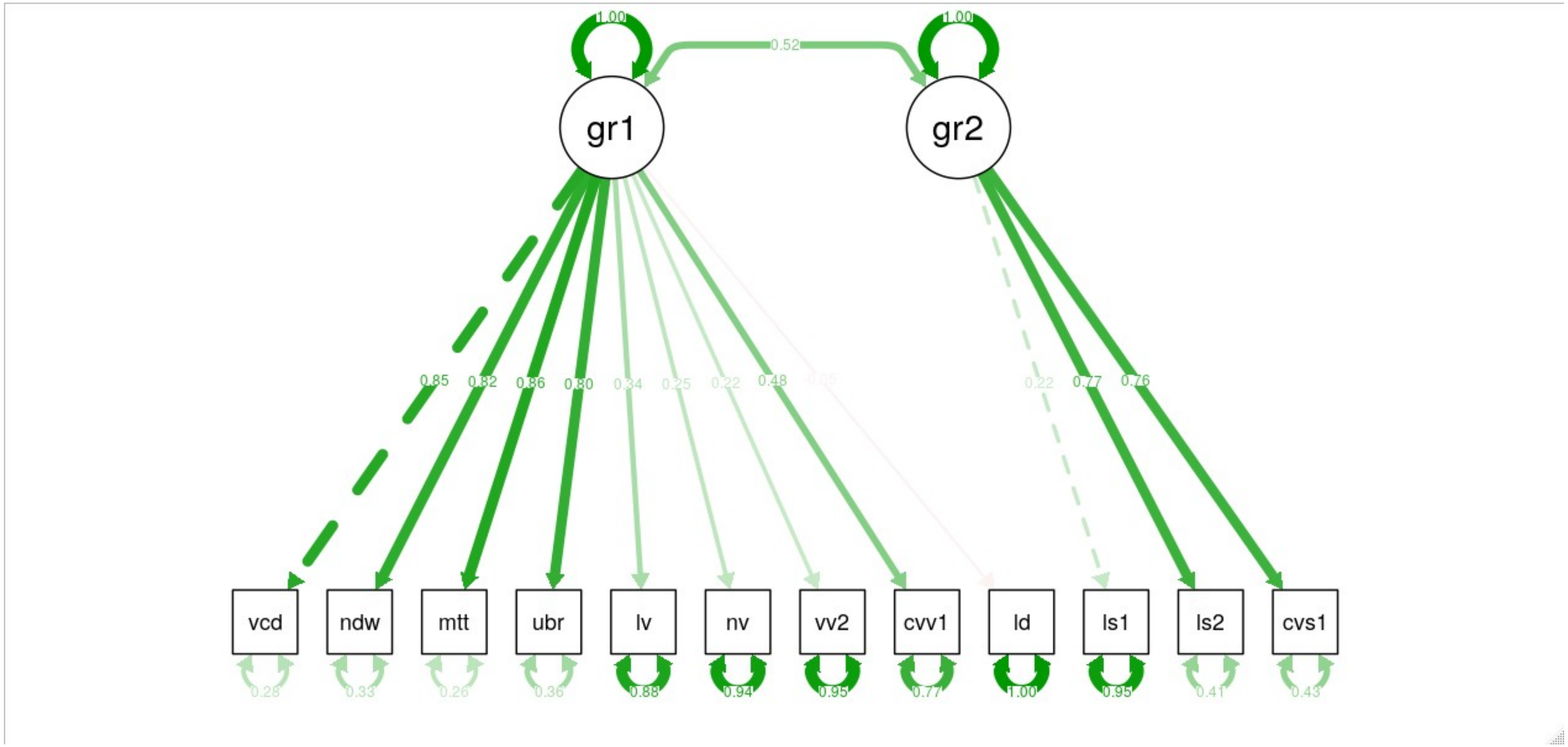
– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are mattr (1), mtld (.96), vocd (.77), ndwerz (.71), ndwesz (.90), msttr (1), hdd (.86), logttr (.59), rttr (.66), uber (.71), maas (-.80), lv (.27), nv (.19), vv1 (.15), vv2 (.11), cvv1 (.35), adjv (.23), ld (-.0), ls1 (1), ls2 (.57), vs2 (.97), cvs1 (1).

Graph 6.6. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model A1



– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are mtld (.96), ndwesz (.89), msttr (.99), hdd (.86), logttr (.61), rttr (.69), maas (-.81), vv1 (.16), adjv (.24), ls1 (.25), ls2 (1), vs2 (.38)

Graph 6.7. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model A2



– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are vocd (.85), ndwerz (.82), mattr (.86), uber (.80), lv (.34), nv (.25), vv2 (.22), cvv1 (.48), ld (-.05), ls1 (.22), ls2 (.77), cvs1 (.76)

The lexical density index (ld) is shown to be the weakest indicator in this factor with a negative factor loading (-0.0) which suggests that it does not belong to the construct of lexical diversity as it was expected and explained in this model's specification. Verb-based measures of vs2 and cvs1 also better represent the construct of lexical sophistication than ls1 and ls2 measures in factor 2 (labelled as 'gr2').

In the diagram of model A1 where the highest-correlated indices are dropped, the remaining indicators in factor 1 better represent the construct of lexical diversity. This is somewhat correct for its parallel model, model A2. However, in both parallel models, the lexical variation indices based on the TTR of word classes (vv1, adjv, lv, nv, vv2) still show the weakest factor loadings with their respective lexical diversity construct, labelled as 'gr1' in both diagrams. In model A2, lexical density is shown with a faint red arrow with the factor loading of -0.05. This suggests that in this model specification the ld measure again shows more distinct characteristics than lexical diversity.

Model B and its two sub-models B1 and B2, however, received better fit indices compared to the type A models. Each of the three B models specifies a fine-grained version of A models, where lexical diversity of word strings, logarithm-based, and TTR-based measures of word classes are assigned separate sub-constructs. Graphs 6.8, 6.9, and 6.10 demonstrate the structure of each group/factor with factor loadings. The graph for model B1 shows a relatively better structure. Nevertheless, these models did not show overall acceptable values for the fit indices, either. However, as table 6.29 shows, among these three models, model B1 resulted in slightly better fit indices of Chi-square, RMSEA, SRMR, and TLI. Section 6.5.2 explores the lexical dataset further to find out a more representative model which can better explain the variations and the relationships between the lexical measures and constructs.

The lexical type B models as demonstrated in above tables (with better fit indices) and graphs (with larger factor loadings) have better fit indices than the type A models. This suggests that separating the lexical diversity indices into sub-constructs results in a better model whereby the lexical measures/indicators are shown to be better representatives of their latent factors.

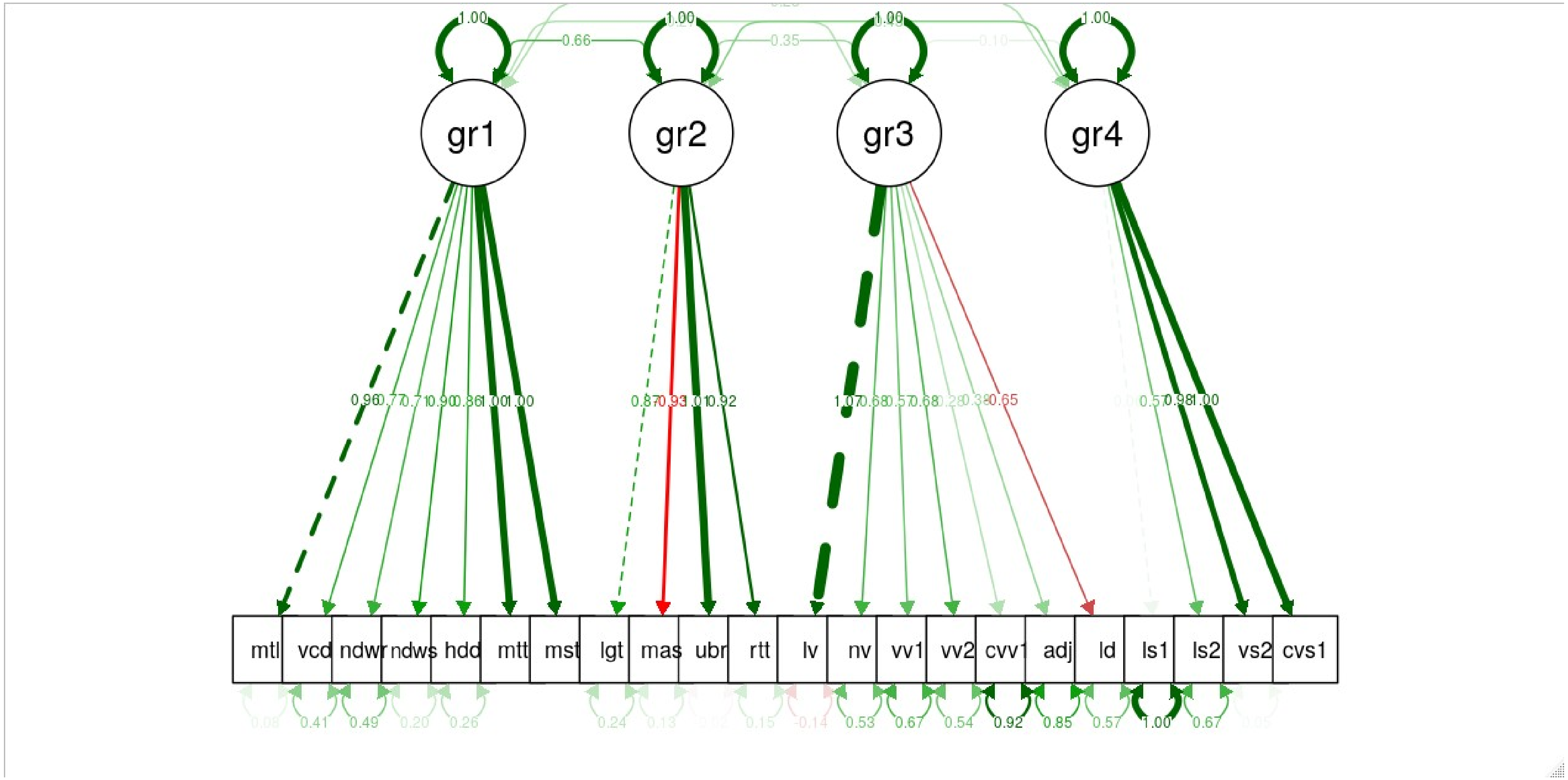
Model B1 receives very high factor loadings for the lexical diversity sub-construct of word strings/segments with mtld (measure of textual lexical diversity), hdd (hypergeometric variant od the D measure), ndwesz (number of different words, first type) and msttr (mean segmental type-token ratio). The same pattern can be seen in model B2 for the same sub-construct of lexical diversity with ndwerz (number of different words, second type), vocd (the original D index), and mattr (moving-average type-token ratio) which also received very high factor loadings. This sub-construct in both models B1 and B2 also received relatively better factor loadings compared to the rest of lexical variables in other latent factors. In the logarithm-based sub-construct of lexical diversity in the parallel models of B1 and B2, maas and rttr indices received better factor loadings which are indicative of indices that better represent that sub-construct. However, rttr (root of type-token ratio) is not a logarithm-based measure, per se, and has been allocated this space only because of its positive correlation with other logarithm-based measures. Taking the squared root of tokens in the rttr index thus is more effective than using logarithm, e.g., for the measures like maas, uber, and logttr. This point will be further revisited in section 6.6 and the discussions of mixed-effects models.

The indices of ls2 (lexical sophistication type two) and vs2 (verb sophistication type two) also better represent the construct of lexical sophistication than the other two indices of ls1 and cvs1 (corrected verb sophistication) as is shown in both B parallel models.

In summary, these findings indicate that even though model B types received better values of fit indices, none of the six lexical models specified earlier, i.e., the two main lexical models of A and B with their respective parallel sub-models, is an optimal lexical model regarding the exact indices that represent the constructs and sub-constructs of lexical complexity and the relationship between and among measures and constructs based on the postgraduate academic writing of English L1, EFL and ESL students. This also shows that there could be a mismatch between the quantification methods of various lexical complexity measures and the conceptual understanding of their assigned constructs proposed based on linguistic classifications. It could be the case that some of these lexical complexity measures may represent/belong to more than one construct and some of them may not represent the construct they are assigned to according to the proposed structure in previous scholarship.

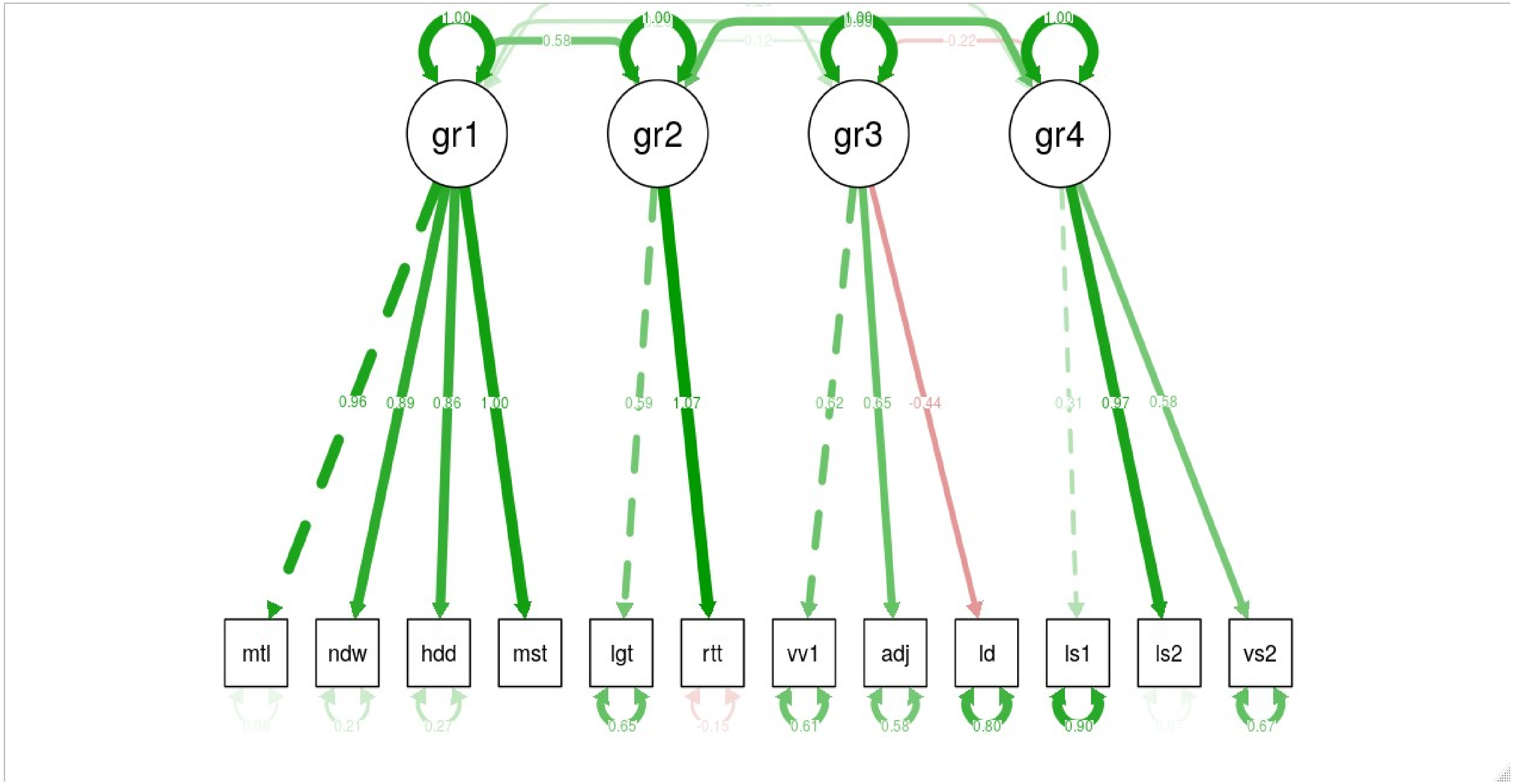
This necessitates the exploration part in McArdle’s recommendation of ‘confirm first, explore second’ to further investigate any possible latent factor (e.g., the constructs and sub-constructs of lexical complexity) that has not been specified in the proposed structure of these indices in the literature as explained earlier in section 6.5. The following exploratory factor analyses also examine which lexical indices are more representative of these lexical constructs regarding this study’s specialised academic writing corpus.

Graph 6.8. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model B



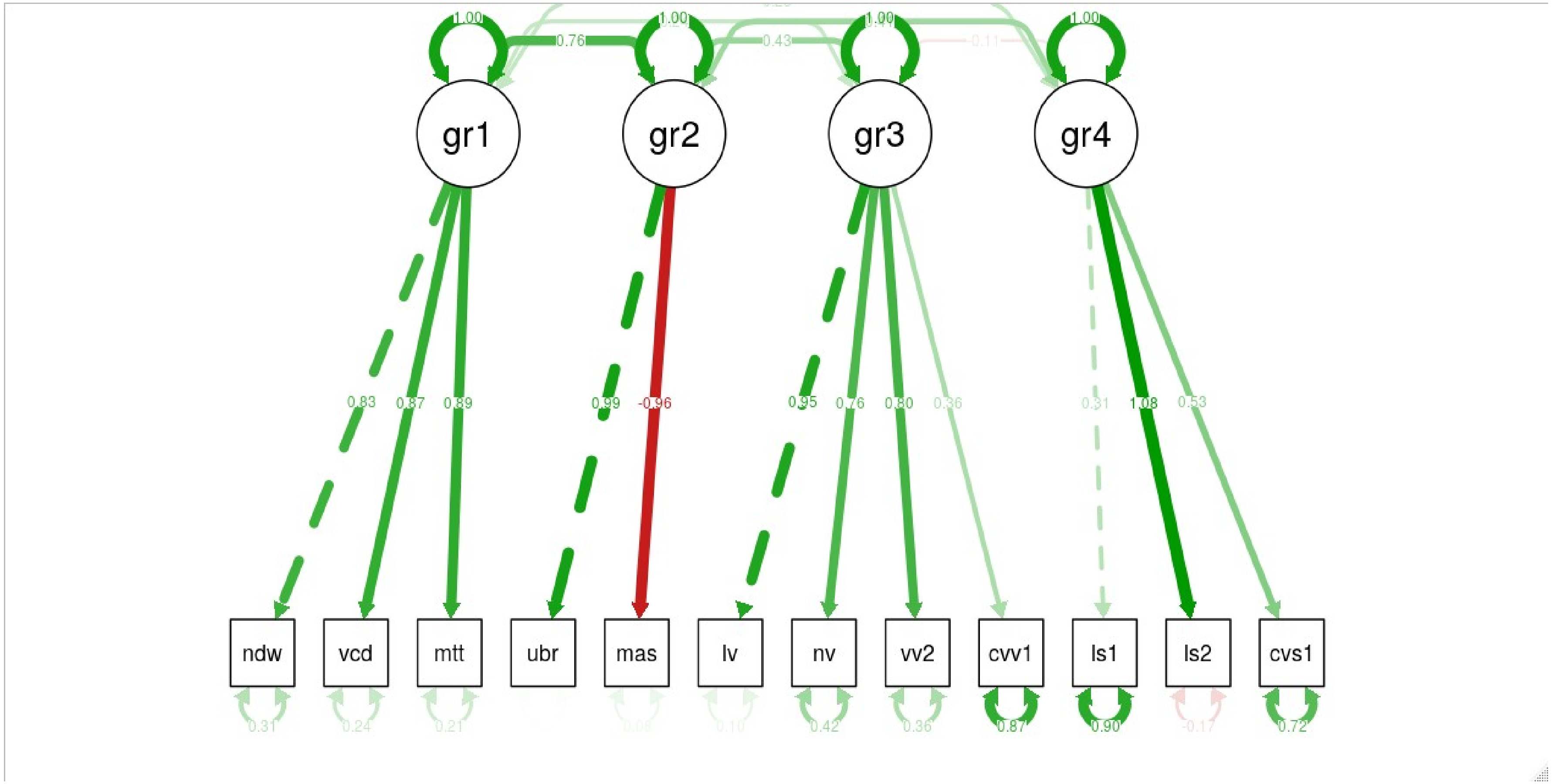
– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are mtld (.96), vocd (.77), ndwerz (.71), ndwesz (.90), hdd (.86), mattr (1), msttr (1), logttr (.87), maas (-.93), uber (1), rttr (.92), lv (1), nv (.68), vv1 (.57), vv2 (.68), cvv1 (.28), adjv (.38), ld (-.65), ls1 (1), ls2 (.57), vs2 (.98), cvs1 (1)

Graph 6.9. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model B1



– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are mtld (.96), ndwesz (.89), hdd (.86), msttr (1), logttr (.59), rttr (1), vv1 (.62), adjv (.65), ld (-.44), ls1 (.31), ls2 (.97), vs2 (.58)

Graph 6.10. Diagram of the structure of lexical constructs and measures with factor loadings for the lexical model B2



– The diagram automatically shortens the lexical labels; labels from left to right with standardised factor loadings are ndwerz (.83), vocd (.87), mattr (.89), uber (.99), maas (-.96), lv (.95), nv (.76), vv2 (.80), cvv1 (.36), ls1 (.31), ls2 (1), cvs1 (.53)

6.5.2. Exploratory Factor Analysis on the Lexical Dataset

A series of exploratory factor analysis (EFA) on the lexical dataset were conducted at this phase to explore the structure of the data regarding the number of factors/constructs and their associated measures using the specifications in 6.5. The number of factors was incrementally increased from the initial four factors until the seven-factor test where no Heywood case nor factor loading greater than one was observed. A Heywood case occurs in the EFA tests if there are too many factors are extracted (i.e., no unique variance of any variable remains for the last factor). The result of the eight-factor model was not significantly different from the seven-factor one with regard to the fit indices (for more details see Appendix C1). The results of the EFA test with seven factors are presented in table 6.30 and the model fit indices are presented in table 6.31.

The negative value of the factor loading of maas should be treated as a positive value. The rest of the measures produced positive medium to large factor loadings as indicated in table 6.30. Following the guideline of Kline (2005), Hu and Bentler (1999), and Lai and Green (2016) for interpreting the model fit indices as demonstrated in table 6.31, the small values of RMSR (e.g., below 0.08) and high values of the Tucker-Lewis index denote good-fitting models. The RMSEA index for this model does not particularly show a good fit; this could be due to high correlations among many lexical diversity measures. For a detailed discussion on the effect of correlations between observed values and a model's RMSEA value see Lai and Green (2016). In the following discussions I will mainly focus on the behaviors of the individual measures and constructs as part of the measure-testing process and, therefore, I do not attempt to re-label the factors/groups based on the new findings of the EFA tests. The main goals of both lexical and syntactic EFA tests in this study are 1) finding the measures that align with the conceptual/linguistic classifications of the constructs, and 2) finding the measures with cross-loadings that may represent more than one construct or sub-construct based on this study's specialised academic writing corpus.

A quick look at table 6.30 shows that most of the measures are loaded on a factor representing the construct or sub-category of a construct that was assigned in 5.3.1.1 in this study based on the existing classifications in the literature (see for instance the classification in Lu, 2012). However, some measures were loaded on a factor that was assumed to belong to another lexical construct as was hypothesised based on the results of confirmatory factor analyses.

Table 6.30. Factors and factor loadings extracted from the exploratory factor analysis on the lexical dataset

Extracted Factors	Lexical Measures with Their Factor Loadings					Cross Factor Loadings Larger than 0.3 in Each Group
Group 1	ndwesz 0.83	mattr 0.96	msttr 0.96	mtld 0.81		ndwerz 0.33 maas -0.43
Group 2	vs2 0.94	cvs1 0.96	cvv1 0.67			Ls2 0.39 vv2 0.36
Group 3	vv1 0.97	adjv 0.43				Lv 0.31 cvv1 0.38
Group 4	ls2 0.56	rttr 0.82	logttr 0.58	uber 0.69	maas -0.57	
Group 5	ndwerz 0.50	hdd 0.54	vocd 0.90			
Group 6	lv 0.68	vv2 0.79	nv 0.84			logttr 0.40
Group 7	ld 0.84	ls1 0.45				logttr 0.38

Table 6.31. Fit indices for the seven-factor lexical model obtained from the EFA analysis

Fit Indices\ Model	Lexical Seven-Factor Model
Empirical Chi-Square	37.09 with prob < 1
Likelihood Chi-Square	567.38 with prob < 3.3e-67
Tucker Lewis Index of factoring reliability	0.84
RMSEA and CIs	0.15, [0.13, 0.16]
RMSR	0.02

--The lexical model is based on 210 number of observations.

The indices that were loaded on group 1/factor 1 all belong to the lexical diversity of word strings or word segments (i.e., ndwesz, mattr, msttr, mtld). The ndwerz measure which is assumed to belong to this factor because of its similar calculation to ndwesz, has a cross loading, appearing to belong both to factor 1 with a low factor loading of 0.3 as well to factor 5 with a slightly higher factor loading of 0.5. This finding suggests that ndwesz better represents the underlying factor/sub-construct of lexical diversity of word strings than ndwerz. This could be due to the random sampling of words in ndwerz compared to the sampling of consecutive words in ndwesz. Mtld and mattr which analyse the texts in sequence order are shown to have very high factor loadings on this group. Both mattr and msttr which showed the perfect correlation in the previous section, have the same factor loadings on this group. However, mattr covers the whole text but msttr may discard some parts of the text that are not included in the fixed-size segments. Their perfect correlation and same factor loadings, therefore, suggest that both work equally well in capturing lexical diversity in long academic texts.

Group 2/factor 2 comprises the vs2, cvs1, and cvv1 indices; the first two indices with very high factor loadings already belong to the verb sophistication category of lexical sophistication construct, but the cvv1 index with a medium factor loading is assumed to belong to the lexical diversity of type-token ratio of word classes. The placement of the verb-based cvv1 index alongside other verb-based sophistication measures indicates that there were no significant differences in the number of verb types and sophisticated verb types in the postgraduate academic writing corpus. This group also includes another two measures of ls2 and vv2 which have cross loadings with small/trivial factor loadings on this group and higher loadings on groups 4 and 6. The inclusion of the vv2 index, even though with small loading, alongside other verb-based measures also indicate a similar amount of verbs, verb types, and sophisticated verb types produced by students in this corpus.

In group 3/factor 3 the vv1 and adjv measures are lined up with other two cross-loaded measures of lv and cvv1 nonetheless with very small/trivial factor loadings. The only measure with a large factor loading value is vv1, a verb diversity measure which calculates the ratio of verb types to the total number of verbs. The rest of the factor loadings in this group are small, denoting that the vv1 index measures a unique feature of verbs and is different than calculating sophisticated verbs. It is also better at capturing the verb variation than the cvv1 measure in group 2 that reduces the effects of verb tokens by making a squared root of them. Furthermore, the cvv1 index is loaded on two factors, showing that it does not have a strong definite presence in a sub-construct.

Group 4/factor 4 witnesses the loadings of three logarithm-based lexical diversity measures of logttr, uber, and maas, along with the rttr measure which takes the square root of tokens in (T / √N), and the ls2 measure of lexical sophistication (sophisticated word types). This finding is in line with the results of correlation tests showing the four measures of rttr, logttr, uber, and maas as highly-correlated measures. The highest factor loading is recorded for rttr with 0.8; the rest of indices in this factor have equally medium factor loadings ranging from 0.56 for ls2 to 0.69 for uber. We can see a similar trend in the diagrams of the SEM models on the lexical dataset. These findings collectively suggest that even though the rttr index is not a logarithm-based measure, taking the square root of tokens in its formula seems a more effective way in reducing the inflation of token counts than using logarithm in the formulas of maas, uber, and logttr. The rttr index surprisingly seems to represent this factor better than the log-based measures; however, this measure did not capture significant between-group differences in the aggregated corpus nor in the genre-separated corpora suggesting that the three groups produced similar amounts of types in relation to the square root of tokens in their texts (i.e., the formula for rttr). The placement of ls2 in this factor shows there was not a huge difference between all word types and sophisticated word types in the corpus of this study. Factor 4 has a moderate correlation with factor 1 (group one and group four correlation = 0.5 as shown in table 6.32). Theoretically, this means these two factors have 50% affinity in what the indices measure and what underlying construct they represent, i.e., the construct of lexical diversity with different quantification methods.

Group 5 marks the strong presence of vocd over the hdd measure which are variants of the D measure as described in 5.3.1.1. Vocd also captured more significant between-group differences than hdd in this study. This could be due to random sampling in vocd compared to directly calculating the lexical probabilities in the case of hdd and/or the effect of sample sizes in the formulas of both measures, that are not geared for very long texts as I indicated earlier. The other indicator in this group is the ndwerz with a factor loading of 0.5, which as mentioned earlier, has a cross loading on the first group with a factor loading of 0.3 as well. Besides, unlike ndwerz, the ndwesz index captured between-group differences in the aggregated data which could qualify it as a better measure capturing variations in the data. Factor 5 has the highest correlation with factor 1 (r = 0.66) which means the measures that are loaded on these two factors are somewhat similar in what they measure and the underlying construct they represent.

The results of factor loadings in group 6, however, shows a straightforward trend with three of lexical variation of TTR of word classes lining up with medium to large loading values. The stronger presence is marked for nv with a factor loading of 0.84, followed by the vv2 and lv indices with 0.79 and 0.68 factor loadings respectively.

Table 6.32. The correlation table for the lexical factors in EFA

	Lex factor 1	Lex factor 2	Lex factor 3	Lex factor 4	Lex factor 5	Lex factor 6	Lex factor 7
Lex factor 1	1						
Lex factor 2	0.15	1					
Lex factor 3	0.22	0.18	1				
Lex factor 4	0.51	0.38	0.10	1			
Lex factor 5	0.66	0.20	0.66	0.44	1		
Lex factor 6	0.22	0.02	0.25	0.26	0.07	1	
Lex factor 7	0.04	-0.12	-0.24	0.17	0.09	-0.22	1

The results of the correlation tests also indicate medium correlation values between these three indices (table 6.21). The only cross loading on this factor is the logttr index with a rather small factor loading of 0.4; this can be due to the fact that all measures in this factor are variants of the T/N ratio and the logttr index is also calculating the log T/ log N. Logttr is also the only measure which loaded on three different factors (i.e., groups 4, 6, and 7), indicating a weaker representation as an indicator for any of these factors. The results of between-group differences on the genre-aggregated as well as genre-separated datasets also do not include the logttr measure as an influential measure which can capture group differences. This cross loading might be a reason for the logttr to explain a large amount of variation (nearly 60 %) in the lexical data as will be discussed in the results of mixed-effects modelling in 6.6.1.

The final factor extracted from the EF analysis lines up two measures of ld and ls1 with 0.8 and 0.4 factor loadings respectively. As mentioned in the previous paragraph, logttr is also cross-loaded on this factor with a small/trivial factor loading of 0.3. Table 6.30 clearly demonstrates that the only noteworthy value in this group belongs to the ld measure, indicating a distinct construct. The highest factor correlation for the lexical data is between groups 5 and 3 as well as between groups 5 and 1 both with r = 0.66; the lowest factor correlation is recorded as r = 0.02 for groups 6 and 2.

6.5.3. Confirmatory Factor Analysis of the Syntactic Data

The same procedures and principles that were outlined in 6.5 in general and in 6.5.1 regarding the CFA models for the lexical data are also followed for the syntactic dataset on the entire corpus. That is, first a general syntactic model based on the classification of Lu and Ai (2015) will be specified to examine the extent to which the structure of the syntactic constructs/factors and their associated measures approximate the structure of these syntactic constructs and measures specified in Lu and Ai (2015) as the main existing classification in the literature. They reported five dimensions in their classification of syntactic complexity measures; however, in this thesis, three measures and one dimension were omitted at the initial measure-selection stage (see section 5.3.1.2), and therefore, the main syntactic model is specified with four factors. To rule out the possible influence of high correlations among the measures (of different constructs) on the values of fit indices of each model, I also ran two parallel models, syn 1 and syn 2, where each sub-model excludes the highest-correlated variables. These syntactic models are specified in diagram 6.3. As with the lexical analysis, the focus of this section will also be on the behaviours of individual measures and constructs as part of the measure-testing process.

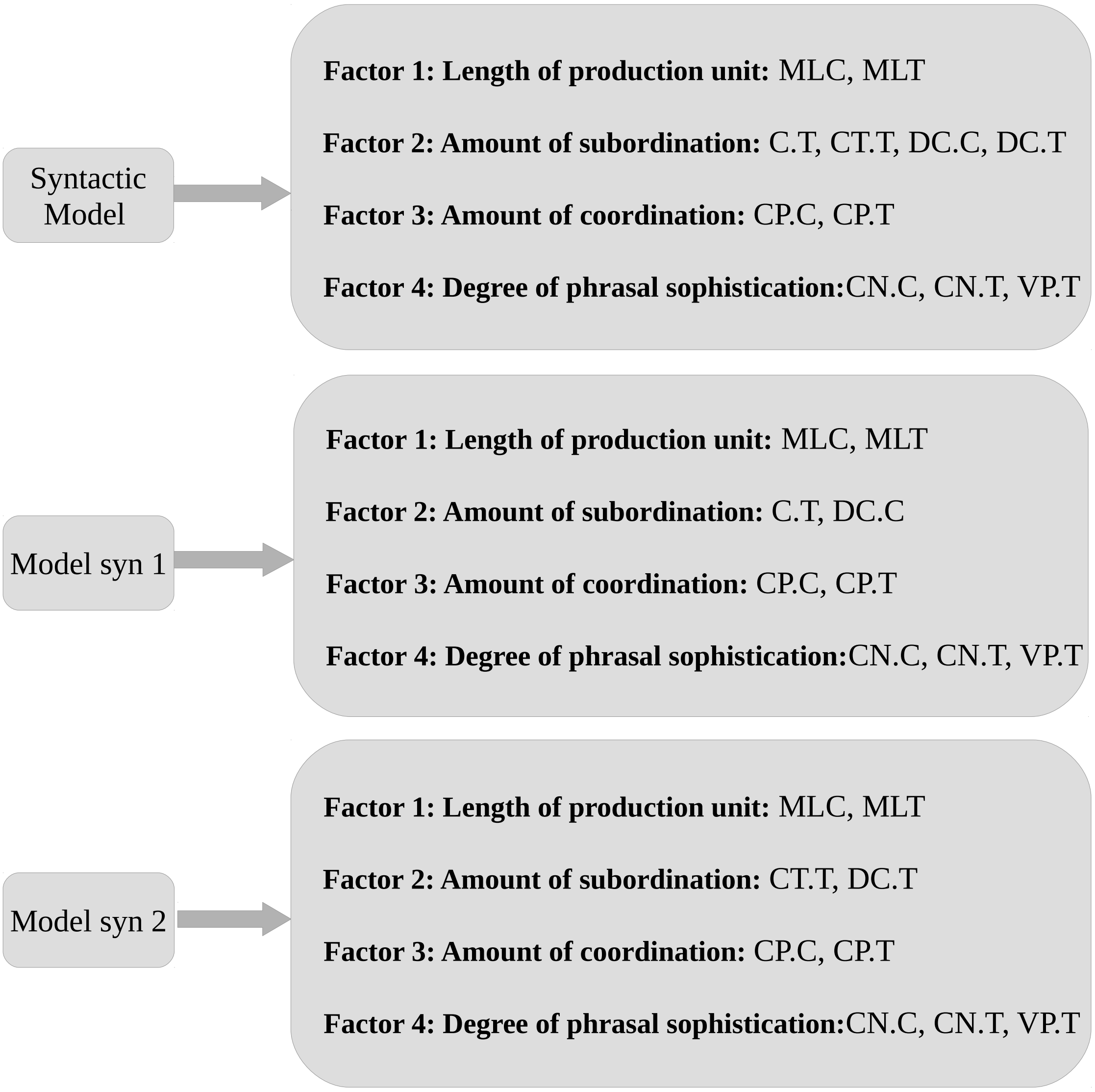
Both syntactic parallel models have the same number of factors as the original model; each parallel model has nine indicators. The only difference between the two sub-models of syn1 and syn2 is the specification of the second factor/group of subordination indices; each sub-model excludes highest-correlated pair of subordination indices. The tests use the robust statistic of Satorra-Bentler which controls any deviations from non-normal distributions in the data. The results of the CFA tests of these three syntactic models are presented and the values of models' goodness-of-fit indices are presented in table 6.33.

Table 6.33. The comparison of syntactic models' fit indices in confirmatory factor analysis

Syntactic Models	**No. Free Parameters**	**Model Chi-Square (p-value)**	**df**	**RMSEA [CIs]**	**SRMR**	**Tucker-lewis Index/ TLI**
Main syntactic Model	28	2512.371 (0.000)	38	0.55 [0.53, 0.57]	0.26	0.44
Model syn1	24	960.347 (0.000)	21	0.46 [0.43, 0.48]	0.38	0.66
Model syn2	24	1860.194 (0.000)	21	0.64 [0.62, 0.67]	0.31	0.30

--All syntactic models are based on 210 number of observations.

Diagram 6.3. Syntactic model and its two sub-models syn 1 and syn 2 with omitted highest-correlated subordination variables



Syntactic Model

Factor 1: Length of production unit: MLC, MLT

Factor 2: Amount of subordination: C.T, CT.T, DC.C, DC.T

Factor 3: Amount of coordination: CP.C, CP.T

Factor 4: Degree of phrasal sophistication:CN.C, CN.T, VP.T

Model syn 1

Factor 1: Length of production unit: MLC, MLT

Factor 2: Amount of subordination: C.T, DC.C

Factor 3: Amount of coordination: CP.C, CP.T

Factor 4: Degree of phrasal sophistication:CN.C, CN.T, VP.T

Model syn 2

Factor 1: Length of production unit: MLC, MLT

Factor 2: Amount of subordination: CT.T, DC.T

Factor 3: Amount of coordination: CP.C, CP.T

Factor 4: Degree of phrasal sophistication:CN.C, CN.T, VP.T

A glance at table 6.33 clearly shows that none of the specified syntactic models is a reliable model due to the unacceptable values of fit indices of Chi-square, RMSEA, SRMR, and the TLI reliability index. However, among the main syntactic model and the syn 1 and syn 2 models, it is the syn1 model which produces slightly better results than the other two with regard to the model performance indices of RMSEA and TLI. Section 6.5.4 and its corresponding table 6.34 as the results of explanatory factor analysis (presented in the next section) on the syntactic dataset shows that the CT/T and DC/T indices which are included in the syn2 model in the CFA analysis, have higher factor loadings and have higher correlation values (table 6.23, r = 0.95). This could be a possible reason for the syn2 model to obtain worse values of fit indices compared to the syn1 model because of the correlation of two subordination indices (table 6.23, r = 0.91) as well as slightly-lower factor loadings in the EFA analysis as will be presented in the next section.

Due to unacceptable fit indices in table 6.33, the syntactic diagrams will not be further explored. Reaching a definite conclusion regarding the syntactic models is not easy due to an overall high correlation of most indices as indicated in table 6.23 Therefore, I will rely on the results of explanatory factor analysis (as will be discussed in section 6.5.4) as well as the clear boundaries between the four main syntactic constructs in the correlation table 6.23, to discuss the overall structure of the syntactic constructs as well as the measures that are classified in Lu and Ai (2015) to quantify them. Further investigation may be required to confirm this current model on larger and more varied academic writing corpora.

6.5.4. Exploratory Factor Analysis on the Syntactic Dataset

The same procedures and principles that were outlined in 6.5 in general and in 6.5.2 for carrying out exploratory factor analyses on the lexical dataset are also followed for the EFA tests on the syntactic dataset on the entire corpus. The parallel analysis of the syntactic data suggested starting the exploratory factor analysis with two factors. The number of factors was then incrementally increased until no Heywood case nor warning messages appeared with the four-factor model. The five-factor model did not converge. Therefore, the four-factor model presented in table 6.34 is chosen as the best model which distinguishes four constructs in a quite similar fashion to the original classification by Lu (2010) and Lu and Ai (2015) with a few exceptions. These results will be later compared with the findings of Yoon (2017) as well.

Table 6.34. Factors and factor loadings extracted from the exploratory factor analysis on the syntactic dataset

Extracted Factors	**Syntactic Measures with Their Factor Loadings**						**Cross Factor Loadings Larger than 0.3 in Each Group**
Group 1	CN/C 0.89			CN/T 0.76			MLC 0.36
Group 2	MLT 0.68	C/T 0.98	CT/T 0.95	DC/C 0.93	DC/T 0.97	VP/T 0.93	CN/T 0.44
Group 3	CP/C 0.97			CP/T 0.96			
Group 4	MLC 0.62						MLT 0.34

The group/factor 1 extracted two measures of CN/C (complex nominals per clause) and CN/T (complex nominals per T-unit) both belonging to the phrasal sophistication category of syntactic structures as specified in 5.3.1.2. The largest factor loading value between the two is recorded for the CN/C measure with 0.89 suggesting that calculating complex nominals per clause vs. per T-unit captures more variance in the data. Lu and Ai's (2015) classification, however, includes the VP/T (verb phrases per T-unit) as another measure in the construct of phrasal sophistication, but in this study, the VP/T index is loaded with the rest of subordination indices in group 2. As I explain in the following paragraph, the second factor's indices are mainly based on T-units. One explanation could be that, overall, the T-units contained similar amounts of verb phrases and dependent clauses, for instance, compared to the number of complex nominals. That is, in the genre- and group-aggregated academic writing corpus of this study, students generally used similar amounts of complex nominals (with categories specified in 5.3.1.2) and subordination structures. Lu and Ai's (2015) results also indicate that overall, English texts from advanced proficiency groups exhibit similar numbers of complex nominals and subordination structures. The first factor also witnesses the presence of a trivial cross-loaded measure of MLC with 0.36 which also reflects sub-clausal complexification such as phrasal structures. As will be discussed below, MLC is moderately loaded as the primary measure of the fourth factor.

Table 6.35. Fit indices for the four-factor syntactic model obtained from the EFA analysis

Fit Indices\ Model	**Syntactic Four-Factor Model**
Empirical Chi-Square	5.27 with prob < 1
Likelihood Chi-Square	868.61 with prob < 1.1e-173
Tucker Lewis Index of factoring reliability	0.55
RMSEA and Cis	0.4, [0.4, 0.5]
RMSR	0.02

– The syntactic model in the EFA test is based on 210 number of observations.

The second factor as presented in table 6.34 includes the largest number of syntactic structures, four of which are subordination indices, one the VP/T index as stated before, and the MLT index of the length of production units. The latter two indices are not in the same construct based on Lu and Ai's (2015) classification. The highest factor loading in group 2 is noticed for the C/T measure (Clauses per T-unit) with 0.98, followed closely by the DC/T measure (dependent clauses per T-unit) with 0.97 which suggests that this subordination construct is strongly represented by the number of clauses produced in the texts. Complex T-units as calculated by the CT/T index is the third strongest indicator in this group with a factor loading of 0.95. The next two large values of factor loadings are also recorded for clause-based structures of DC/C and VP/T which measure dependent clauses per clause and verb phrases per T-unit, both with a factor loading of 0.93. The inclusion of verb phrases in this category of subordination indices is an unexpected result that I discussed earlier. The other seemingly misplaced index, MLT, though has only a medium factor loading compared to other values in this group. The placement of MLT which is assumed to line up with MLC in the length of production units, along other subordinating indices based on the classification of Lu and Ai (2015), could be due to the fact that MLT in the first place calculates the main clause with any subordinate clauses embedded in it as explained in section 5.3.1.2. We can see a similar result as reported in Yoon (2017) where MLT is loaded together with clause-level syntactic measures such as C/T, and not with other length-based measures like MLC. As will be discussed regarding the factor loadings on group four in this study (see table 6.34) MLT has cross loadings on both length-based and subordination dimensions, making it a relatively weak indicator of either of the constructs. This group also cross-loaded the CN/T measure with a very small factor loading of 0.44; the CN/T measure has a moderate factor loading on the first factor, though.

Table 6.36. The correlation table for the syntactic factors in EFA

	Syn factor 1	Syn factor 2	Syn factor 3	Syn factor 4
Syn factor 1	1			
Syn factor 2	0.23	1		
Syn factor 3	0.46	0.08	1	
Syn factor 4	0.64	-0.01	0.51	1

The results of the third factor, however, is straightforward and consistent with the classification of coordination indices of CP/C and CP/T which calculate the ratio of coordinate phrases per clause and per T-unit. Both measures have very similar factor loadings of 0.97 and 0.96 respectively. This finding is in line with the result of the correlation test which showed a high correlation between these two indices at r = 0.89. However, between the two measures, only CP/C showed a significant between-group difference in the result and discussion section (tables 6.18 and 6.20) at the strict Bonferroni-corrected level of 0.004 for two sets of comparison of ESL-EFL and NS-EFL with the EFL group obtaining larger values (see the discussion in 6.3.1). The CP/C index also captured a larger amount of variation compared to CP/T (table 6.38) as will be discussed in the following section regarding the mixed-effects modelling. These findings collectively suggest that of the two, the CP/C index would be a better choice for future researchers studying syntactic structures in academic writing. Unlike this study and Lu and Ai's (2015) classification, in Yoon's (2017) factor analysis, the CP/C index is loaded with the CN/C (complex nominals per clause) index, suggesting that coordination and phrasal-based syntactic measures could also be affected by the variations in the corpus as I discuss below.

The final factor extracted only one measure of MLC (mean length of clause) which calculates the mean length of clauses in the texts; the MLT (mean length of T-unit) index is only cross-loaded with a very small factor loading of 0.34. As explained earlier, the existing theoretical classification in the literature includes both MLC and MLT as a construct which analyses the mean length of the production unit. The trivial factor loading of MLT indicates that MLC better represents the mean length of production unit construct and better explains the variation in this study's academic writing corpus. Since the loading of MLT along with the second group is only marked with a medium value of factor loading, further investigation in academic writing corpora is called for to examine whether the two indices, in fact, belong to the same theoretical construct. Yoon (2017, p. 137) as discussed earlier, also found that MLC and MLT loaded on different constructs suggesting that they may be tapping "different levels of syntactic complexity depending on its base production unit". However, unlike this study, in Yoon's factor analysis, MLC loaded with phrasal level measures such as CP/C (coordinate phrases per clause) and CN/C (complex nominals per clause), suggesting that MLC values might heavily depend on the variations in the corpus in terms of topic, writers' English language backgrounds, text length, corpus size, etc. Yoon's (2017) study investigated a corpus of written argumentative essays written by college-level Chinese EFL learners on two general topics, with essay length ranging from 221 to 250 words (total size of the corpus is 280, 203).

The highest factor correlation for the syntactic data (table 6.36) is marked as r = 0.64 for groups 4 and 1; the lowest factor loading is between groups 4 and 2 with r = -0.01.

6.5.5. A Summary of Key Findings of 6.5

This section mainly focused on the measure-testing process of the individual measures and their representative constructs. Confirmatory factor analyses in this section were conducted to examine the extent to which the classification of the lexical and syntactic complexity constructs and their representative measures as defined by the linguists can be supported by the data from a postgraduate academic writing corpus. The findings show that none of the lexical and syntactic models nor their sub-models produced acceptable fit indices in the confirmatory factor analyses; this suggests that the structures of lexical and syntactic indices analysed in the postgraduate academic writing corpus are not completely consistent with the proposed structures/classifications suggested in the literature. Even though the effect of high correlations among various measures (e.g., the measures that represent different sub/constructs) was controlled to some extent by way of parallel modelling, the remaining moderate correlations among the measures in each dataset could have contributed to unacceptable fit indices. Despite these negative findings, some interesting patterns were observed especially regarding the comparison of the main models with sub-models. For instance, model fit indices and the diagrams generated by the lexical models show that a better structure for the lexical diversity measures is obtained when they are assigned various sub-constructs based on their quantification methods (e.g., the classification in this study in section 5.3.1.1) rather than taking all lexical diversity measures under one category (e.g., table 2.1 based on all lexical diversity measures proposed in the literature).

In the next step, a series of exploratory factor analyses were administered to explore this discipline-specific postgraduate academic writing corpus further and to compare the obtained structure with the existing classifications of lexical and syntactic complexity. The explanatory factor analysis on the lexical data produced a seven-factor model with fine-grained categories of lexical diversity; four misplaced indices did not line up on the expected factors based on the classifications in the literature (e.g., Lu, 2012). However, the EFA results on the syntactic data showed an overall consistency with the categories of syntactic constructs and the corresponding indices proposed by Lu and Ai (2015) except two misplaced measures. The results of the exploratory factor analysis on both lexical and syntactic datasets also show that most measures are loaded based on the calculation methods of the indices rather than purely conceptual basis (i.e., the linguistic categories).

Regarding some specific findings in the lexical dataset about the distinctiveness of the constructs, the ld measure is shown to be the sole indicator of a distinct construct of lexical density. Lexical sophistication indices did not all group together in a distinct construct.

Lexical diversity as a construct is best explained by various sub-constructs based on the calculations methods rather than the conceptual/linguistic categories. The ls1 and ls2 indices of lexical sophistication did not cluster with each other nor with the other two verb-based sophistication measures; both lexical sophistication measures received low to moderate factor loadings only indicating low overall variation in the aggregated data. The vv1 index could better capture verb variation than cvv1 and vv2 indices which calculate verb types with different denominators. Vv1 as the ratio of verb types to verb tokens is also shown to be a distinct measure dominating a separate lexical construct. Among the lexical diversity measures, logttr has received mixed results so far: it explains high (nearly 60 per cent of the) variation in the lexical data among other indices (table 6.37 in the next section) while representing a weak indicator (table 6.30) due to cross loading on three factors, and capturing between-group variations mostly at the 0.05 level on genre-separated datasets (tables 6.7 to 6.13).

Finally, the EFA tests result in some specific findings in the syntactic dataset. For instance, the MLT (mean length of T-unit) index received very small factor loading in the construct measuring the length of production units (group 4) but received moderate factor loading in the subordination construct (group 2) contrary to the expected result based on the proposed classification of Lu and Ai (2015). This could be due to the fact that MLT calculates the main clause and any subordinate clauses embedded in it (see for instance section 5.3.1.2). The VP/T measure which calculates the number of verb phrases in T-units is another misplaced measure according to the Lu and Ai (2015) classification. This measure has a very high factor loading along with other subordinating indices in group 2.

In summary, while most of these indices did line up on their expected factors/constructs as specified in the measure-selection process, several measures lined up on different constructs or sub-constructs, mainly due to similar quantification methods. Linguistically, this suggests that the type of corpus, e.g., a specialised academic corpus in this study vs a general English writing corpus in previous works, affects the structure/classification of these complexity measures and what we conceptually assign as, for example, lexical density vs diversity vs sophistication. This is because in chapter two I have already discussed the distinctiveness of these constructs from theoretical and linguistic points of view, using various examples as well as previous scholarship. The EFA results, however, show that lexical density (ld) and a measure of lexical sophistication (ls1) are lined up on the same factor indicating that in a specialised academic writing corpus such as the dissertations in this study, we can find similar numbers of lexical items and sophisticated lexical items which is suggestive of overall sophisticated nature of the texts (e.g., discipline-specific terminologies). The same goes for the two sophisticated verb-based measures (vs2 and cvs1) and a diversity verb-based measure (cvv1) that are grouped in the same factor. This result confirms that there are similar numbers of verb types and sophisticated verb types in this specialised academic writing corpus indicating that, all students irrespective of their English language backgrounds, have used a much higher rate of sophisticated words/specialised words that we would usually find in a general English language corpus. The linguistic interpretation of the syntactic EFA results is already integrated within the discussion of the measure-testing process. The results, for example, show that apart from measuring the length of clauses, MLC is also indicative of phrasal complexity, and therefore, can be considered as a multi-trait index. The same goes for verb phrases and mean length of T-units. The construct of coordination, on the other hand, seems to be distinctly different from other syntactic constructs.

The type and number of cross loadings in this study are also important as they show the measures that do not uniquely quantify/represent a specific construct but rather are representative of more than one linguistic/conceptual construct based on a specialised academic writing corpus. These cross loading findings are important for academic writing research in different ways: researchers who want to restrict the number of indices which are generally indicative of linguistic complexity (rather than distinct constructs) may use these multi-trait indices, and researchers who need distinct measures representative of a certain construct may avoid these measures with cross loadings.

In the next section, I will continue this measure-testing process and the behaviours of the individual measures but in respect to a model selection approach and how the variation in the values of these complexity measures can be affected by text-extrinsic and text-intrinsic factors.

6.6. Identifying the Best-fitting model with Linear Mixed-effects Modelling: An Explanation of the Variations in the Data and the Effects of Groups and Rhetorical Predictors

In the previous sections, the relationships between and among lexical and syntactic complexity indices and constructs as well as their overall and specific structures and models were demonstrated. In this section, I will elucidate the effects of each rhetorical section and the groups of students on the lexical and syntactic values. This is obtained by the second type of statistical modelling in this research, linear mixed-effects modelling. The rationale behind a mixed model analysis was to understand which of the ‘rhetorical section’ and ‘groups’ as the predictors of students’ performance indicated by the lexical and syntactic values could explain more variation in each measure’s estimates and whether taking both explanatory variables vs. only one variable into the models could predict larger values (e.g., higher lexical and syntactic complexity of the texts). Furthermore, the linear mixed-effects models as opposed to the general linear models, incorporate by-subject variability and account for multiple responses/data points for each student for each rhetorical section of the dissertations (and to fulfil the assumption of the independence of data points).

As with other inferential statistics, the mixed-effects test also requires certain statistical assumptions to be met for the validity of the results. The Variance Inflation Factors (VIF) indicate only mild collinearity between predictors (i.e., between groups and rhetorical sections labelled as ‘genre’; the VIF < 4 results in the link provided in Appendix B) and the visual inspection of residuals plots did not indicate any severe deviations from normality and homoscedasticity as demonstrated in graphs 6.2 and 6.4. However, I employed a robust statistic with bootstrapping which compensates for any deviations from normality in the residuals, e.g., the upper right data points in the syntactic graph 6.4. To counterbalance the effect of multiple significance testing (e.g., for several lexical and syntactic indices) and the possible increase of type I error rates, the new Bonferroni corrected alpha levels of 0.002 for the lexical and 0.004 for syntactic tests were applied to the models’ p-values.

The mixed-effects modelling was conducted using the *lme4* package (version 1.1-21; Bates et al., 2015). R^2 values are the R^2m output from the *MuMIn* package (version 1.43.6; Bartoń, 2019) which shows the R-squared for the fixed effects in a model. For an in-depth discussion of how this type of marginal R^2 represents the variance explained by the fixed effects see Nakagawa and Schielzeth (2013). The model comparison fit indices of AIC (Akaike Information Criterion; smaller AIC values indicate better model fit), Log-Likelihood (the likelihood of particular values for model estimates or parameters; larger values denote better model fit given the data), and Chi-square and its accompanied p-values from the likelihood ratio tests of the nested models were derived from the *anova* function in Base R. AIC and log-likelihood values can be positive or negative. For model selection, the model which has the smallest AIC and the largest log-likelihood and R-squared values is selected as a minimal adequate model, i.e., a model which explains the maximum variance with the minimum number of predictors. In contradictory findings, all three fit indices are taken into account for explaining the models. This is because AIC puts a penalty on complex models and is better suited for prediction purposes, e.g., with new data, while R-squared explains the variation in observed data, and Log-likelihood is used often in significance testing. The *bootMer* function from the lme4 package and *boot.ci* function from the *boot* package were used to get the bootstrapped estimates and confidence intervals of the fixed effects, and the *Pbmodcomp* function from the *pbkrtest* package (version 0.4-7; Halekoh & Højsgaard, 2014) generated bootstrapped p-values for the models' comparisons to compensate for any deviations from normality and homoscedasticity in the residuals. Complementary results of the fixed effects are provided via the link in Appendix B.

For each measure four models were specified: two models for investigating the separate effects of the rhetorical sections (labelled as 'genre' in the tables) and the groups of students, one model which incorporates both predictors as separate fixed effects, and one model which checks the interaction of these two variables. The interaction effect examines whether the relationship between the first IV (e.g., groups) and DV (a measure) is different at different levels of the other IV (e.g., different rhetorical sections). All four models have the exact same random effect: a random intercept to account for the by-student variability (the models with the random slope of rhetorical sections' variation led to convergence issues due to overparametrisation and hence could not be included). All models use only one random intercept for the effect of students (i.e., to account for the variability among the six data points produced by each student for each rhetorical section). Tables 6.37 (in section 6.6.1) and 6.38 (in section 6.6.2) present model comparisons with the mentioned fit indices.

6.6.1. Linear Mixed-effects Lexical Models

In this section the effects of four types of mixed-effects models on the values of 22 lexical complexity measures will be examined. The new Bonferroni-corrected alpha level of 0.002 for the lexical models is applied to the p-values (indicated by values in bold font in the table) to compensate for the possible false-positive findings. The bootstrapped p-values in the last column accompany the Chi-square results of the likelihood ratio tests for nested models and indicate whether the difference between a model and its previous model in the column is large enough, i.e., not due to chance. For instance, for the ld measure, the p-value of 0.0001 means that the difference between the fit of 'groups + genre' and 'genre'-only models is large (significant linear effect) and that adding groups to the previous model (a linear combination of the two effects) improves the fit.

Table. 6.37. Model comparison fit indices for the linear mixed-effects lexical models

Lexical Measure	**Model**	**Model Comparison Indices**					
		R2	**AIC**	**LogLik**	**Chisq**	**df-diff**	**Boot p-value**
ld	groups	0.051	-7609	3809			
	genre	0.020	-7642	3829			
	groups+ genre	0.072	-7656	3838	18.31	2	**<.001**
	groups*genre	0.081	-7663	3851	26.49	10	0.0031
ls1	groups	0.016	-954	482			
	genre	0.020	-987	501			
	groups+ genre	0.037	-990	505	6.95	2	0.0308
	groups*genre	0.050	-999	519	29.09	10	**0.0012**
ls2	groups	0.001	-2922	1466			
	genre	0.43	-3881	1948			
	groups+ genre	0.43	-3878	1949	1.02	2	0.59
	groups*genre	0.44	-3888	1964	29.66	10	**<.001**
vs2	groups	0.0001	4158	-2074			
	genre	0.173	3873	-1928			
	groups+ genre	0.173	3876	-1928	0.11	2	0.94
	groups*genre	0.178	3887	-1923	9.40	10	0.49
cvs1	groups	0.0002	1379	-684			
	genre	0.232	991	-487.60			
	groups+ genre	0.232	995	-487.51	0.17	2	0.91
	groups*genre	0.237	1006	-483	8.96	10	0.53
ndwerz	groups	0.01	4825	-2407			
	genre	0.18	4518	-2251			
	groups+ genre	0.19	4509	-2244	13.16	2	**0.001**
	groups*genre	0.21	4496	-2228	33.61	10	**<.001**
ndwesz	groups	0.04	5058	-2524			
	genre	0.06	4957	-2470			
	groups+ genre	0.11	4937	-2458	24.52	2	**<.001**
	groups*genre	0.13	4917	-2438	39.38	10	**<.001**
rttr	groups	0.0003	6068	-3029			
	genre	0.5852	4757	-2370.8			
	groups+ genre	0.5856	4761	-2370.6	0.37	2	0.82
	groups*genre	0.59	4754	-2357	26.25	10	0.003
logttr	groups	0.003	-5025	2517			
	genre	0.574	-6308.4	3162			
	groups+ genre	0.578	-6308.4	3164	3.97	2	0.13
	groups*genre	0.58	-6329	3184	40.77	10	**<.001**
uber	groups	0.01	5314	-2652			
	genre	0.10	5086	-2535			
	groups+ genre	0.12	5083	-2531	7.04	2	0.02
	groups*genre	0.15	5021	-2490	81.35	10	**<.001**

Lexical Measure	**Model**	**Model Comparison Indices**					
		R2	**AIC**	**LogLik**	**Chisq**	**df-diff**	**Boot p-value**
lv	groups	0.002	-265	137			
	genre	0.41	-1090	553			
	groups+ genre	0.421	-1088	554	2.49	2	0.28
	groups*genre	0.425	-1080	560	11.65	10	0.30
vv1	groups	0.002	535	-262			
	genre	0.114	366	-175			
	groups+ genre	0.116	369	-174	1.73	2	0.42
	groups*genre	0.12	381	-170	7.94	10	0.63
cvv1	groups	0.0005	2056	-1023			
	genre	0.3601	1411	-697.5			
	groups+ genre	0.3606	1414	-697.2	0.48	2	0.78
	groups*genre	0.364	1424	-692	10.14	10	0.42
vv2	groups	0.003	-1913	961			
	genre	0.112	-2070	1043			
	groups+ genre	0.115	-2069	1044	2.82	2	0.24
	groups*genre	0.117	-2052	1046	3.49	10	0.96
nv	groups	0.004	76	-33			
	genre	0.320	-440	228			
	groups+ genre	0.324	-441	230	4.66	2	0.09
	groups*genre	0.33	-435	237	14.72	10	0.14
adjv	groups	0.005	-3491	1750			
	genre	0.032	-3530.6	1773			
	groups+ genre	0.038	-3530.8	1775	4.23	2	0.12
	groups*genre	0.04	-3523	1781	13.03	10	0.22
maas	groups	0.02	-9348	4679			
	genre	0.12	-9597	4806			
	groups+ genre	0.15	-9608	4814	14.60	2	**<.001**
	groups*genre	0.19	-9680	4860	92.23	10	**<.001**
mattr	groups	0.07	-5210	2610			
	genre	0.11	-5443	2729			
	groups+ genre	0.19	-5473	2746	34.60	2	**<.001**
	groups*genre	0.21	-5508	2774	54.40	10	**<.001**
msttr	groups	0.07	-5167	2589			
	genre	0.11	-5384	2700			
	groups+ genre	0.19	-5416	2718	36.38	2	**<.001**
	groups*genre	0.21	-5447	2743	51.28	10	**<.001**
hdd	groups	0.02	-5492	2751			
	genre	0.19	-5883	2949			
	groups+ genre	0.22	-5893	2957	14.46	2	**<.001**
	groups*genre	0.25	-5959	2999	85.16	10	**<.001**
mtld	groups	0.06	9332	-4661			
	genre	0.10	9081	-4532			
	groups+ genre	0.17	9058	-4519	27.41	2	**<.001**
	groups*genre	0.18	9033	-4496	44.90	10	**<.001**

Lexical Measure	Model	Model Comparison Indices					
		R2	AIC	LogLik	Chisq	df-diff	Boot p-value
vocd	groups	0.03	10939	-5464			
	genre	0.24	10392	-5188			
	groups+ genre	0.27	10379	-5179	17.19	2	**<.001**
	groups*genre	0.28	10369	-5164	30.10	10	**<.001**

– Df-diff is the difference between the degrees of freedoms of the two models being compared.
– Degrees of freedom for the groups-only model is 5, for the genre-only model is 8, for the groups + genre model is 10, and for groups * genre model is 20.

As explained in 6.6, four models were specified and conducted for each lexical complexity measure to investigate different patterns of the effects of fixed effects (i.e., the study's predictors or explanatory variables) on the variation of values of each index and to find out which model better explains the variability in students' scores with regard to each measure. For each measure, the first model isolates the effect of groups and the second model isolates the effect of the rhetorical sections of dissertations (labelled as 'genre' in the tables) on the variability of indices' values. The third model inspects the additive effects of groups and genre, and the final model explores the interaction effects of these two predictors. The models take the EFL (among the groups) and abstract (among rhetorical sections) as the baseline (see the results in the repository).

Table 6.37 shows that most of the lexical indices' values are most affected by the interaction of groups and rhetorical sections as specified with groups*genre. This evidence is derived from the larger R-squared and Log-likelihood values, and smaller AIC values. These indices include lexical density, the two lexical sophistication indices of ls1 and ls2 as well as the lexical diversity measures based on word-strings (ndwerz, ndwesz, mattr, msttr, hdd, mtld, vocd-D), rttr, and logarithm-based lexical diversity measures (uber, logttr, maas). This means, for example, that the value of lexical density of a particular group also depends on the rhetorical section being analysed (i.e., not all rhetorical sections have the same effect on lexical density). The highest effect of the predictors in this model based on R-squared values is recorded for the rttr (root type-token ratio, R^2 = 0.59) and logttr (logarithm type-token ratio, R^2 = 0.58) measures and the lowest effect of this interaction is spotted for the adjv index (adjective variation, R^2 = 0.04).

Taking smaller AIC values as the model-selection criterion, the 'genre'-only models explain the greatest amount of variation for the measures of vs2, cvs1, lv, vv1, cvv1, and vv2 (mainly verb-based measures), and the 'groups + genre' models explain the greatest amount of variation for the nv and adjv indices. A significant additive model shows that the effect of one factor (e.g., groups) does not depend on the level of the other factor (e.g., various rhetorical sections of dissertations). In the case of these results, it shows, for example, that the production of varied adjectives by different groups does not depend on any specific rhetorical section, i.e., all groups in all rhetorical sections produced similar amounts of adjectives. In the case of verb-based measures and ‘genre’-only models, this can also be seen in the findings in section 6.3.1 where we have witnessed that these lexical variation measures based on word classes as well as these verb-based measures could capture between-group differences for specific rhetorical sections such as literature review and result sections with medium-large effects.

The next-best model for explaining the values of all lexical measures based on smaller AIC and larger R-squared values, is the additive effects of groups and rhetorical sections, (specified in table 6.37 as groups + genre); however, for some of the measures, the genre-only model shows equally good fit or better fit (e.g., for the indices of ls2, vs2, cvs1 which are all lexical sophistication indices), and for a few other measures, the difference between the model fit indices for the two models of genre-only and groups + genre is only marginal (e.g., for the measures of ndwerz (number of different words, second type), rttr, logttr, vv1 (verb variation type one), cvv1 (corrected verb variation of type one), nv (noun variation), and adjv which are all lexical diversity measures). With respect to the differences between the interaction models and the additive models, for most of the lexical measures, the interactional models are significantly better models as indicated by the bootstrapped p-values of the chi-square tests along with other fit indices. These indices include ls1 or first type of lexical sophistication, ls2 or the second type of lexical sophistication, ndwerz that is the second type of number of different words, ndwesz as the first type of number of different words, logttr, uber and maas all three as logarithm-based indices, mattr or the moving-average TTR, msttr or the mean segmental TTR , hdd as the hypergeometric variation of D, mtld or the measure of textual lexical diversity, and vocd or the original D index. For the rest of measures (i.e., ld, vs2, cvs1, rttr, lv, vv1cvv1, vv2, nv, adjv), however, the differences in these two models are not significant, suggesting that the variations in the values of these indices can be best explained both by an additive effect as well as an interaction effect of groups and genre.

Linguistically speaking, these results show that the values of most of lexical complexity measures for each group heavily depend on the rhetorical section of that text. For instance, the amount of sophisticated lexical items produced by each group depends on which rhetorical section is being analysed. In this study, the English L1 and ESL groups in their introduction sections produced more non-repetitious sophisticated words (ls2) than the EFL group compared to other rhetorical sections.

The models with the isolated effect of groups recorded the lowest fit values of R-squared indicating that only identifying different groups of students for a study on lexical complexity, is not the most effective way of finding the variation in these measures' values. The highest R^2 value recorded by the groups-only models, is shown for the mattr and msttr indices, each with the R^2 value of 0.07.

The mixed-effects models' results also support the overall trend that we noticed in the analyses of variance and mean differences of the three grroups. The results of lexical density, for instance, show significant increases from EFL abstracts to ESL abstracts and from EFL abstracts to English L1 abstracts. A similar pattern appears across genre-only effects with regard to the EFL group which is taken as the reference level. The results indicate an overall increase in the lexical density of texts of the EFL group as we proceed from abstracts to the final rhetorical sections in the dissertations except method sections. The interaction effects of groups and rhetorical sections for all levels, on the other hand, do not show a straightforward trend regarding the production of lexically dense texts for every single rhetorical section. But there is an overall incremental increase in the density of lexis (i.e., the density of lexical words) when we read through dissertation sections from EFL abstracts towards conclusion sections of the other two groups, indicating an overall more lexically-dense texts of the ESL and English L1 groups across rhetorical sections. This is despite the fact that introductions of English L1s are significantly lower in lexical density, indicating that English L1 students in this study have started the texts with easier constructions (e.g., including more functions words to clarify the relationships between constructions) and have proceeded with higher density of lexis in the subsequent rhetorical sections. A similar pattern is observed for the ESL group in how densely they packed the lexical items in sytactic constructions starting with lower lexical density in introductions and a general trend of increase in subsequent sections. The sign of these interaction for ESL and English L1 groups is negative indicating that their abstracts are more lexically dense than introduction, etc.

Similar patterns can be observed in most lexical diversity measures as well, especially the lexical diversity measures based on word-strings (mattr, msttr, mtld, hdd, and vocd). That is, overall, the ESL and English L1 groups produced more non-repetitious lexical words across rhetorical sections compared to the EFL students. But this trend is not linear across the sections for ESL and English L1 writers: their abstracts, introductions and conclusions seem to include more diversification of lexis than their literature review, method, and result sections. This pattern is also consistent for lexical sophistication measures across groups and rhetorical sections, except method, result, and conclusion sections in which EFL texts appear to be slightly more complex in terms of the production of less-frequently-used words (especially as calculated via ls1). The link to full results of these mixed-effects models for all complexity measures is provided in Appendix B.

The findings of the model comparison for all lexical measures also indicate that the isolated effect of genre explains more variation in the data than the isolated effect of groups as the main predictors of the models (i.e., the two non-nested models). That is, more variations in the values of these complexity measures can be found among different rhetorical sections compared to groups of students. This is demonstrated in table 6.37 with much smaller AIC values for the genre-only model.

6.6.2. Linear Mixed-effects Syntactic Models

The procedures for mixed-effects modelling on the syntactic dataset is the same as the lexical dataset. The new Bonferroni-corrected alpha level of 0.004 for the syntactic models is applied to the p-values (indicated by values in bold font in the table) to compensate for the possible false-positive findings.

Table 6.38. Model comparison fit indices for the linear mixed-effects syntactic models

Syntactic Measure	**Model**	**Model Comparison Indices**					
		R2	**AIC**	**LogLik**	**Chisq**	**df-diff**	**Boot p-value**
MLT	groups	0.06	6957	-3473			
	genre	0.04	6869	-3426			
	groups+ genre	0.111	6847	-3413	25.54	2	**<.001**
	groups*genre	0.118	6849	-3404	18.39	10	0.04
MLC	groups	0.01	5460	-2725			
	genre	0.11	5235	-2609			
	groups+ genre	0.13	5231	-2605	8.10	2	0.01
	groups*genre	0.14	5233	-2596	18.56	10	0.04
C/T	groups	0.078	148	-69			
	genre	0.075	11.51	2.24			
	groups+ genre	0.15	-19.51	19	35	2	**<.001**
	groups*genre	0.16	-34.12	37	34	10	**<.001**
CT/T	groups	0.10	-1661	835			
	genre	0.06	-1738	876			
	groups+ genre	0.17	-1789	904	55.87	2	**<.001**
	groups*genre	0.19	-1807	923	37.77	10	**<.001**

Syntactic Measure	Model	Model Comparison Indices					
		R2	AIC	LogLik	Chisq	df-diff	Boot p-value
DC/C	groups	0.08	-2730	1370			
	genre	0.09	-2885	1450			
	groups+ genre	0.18	-2926	1473	45.56	2	**<.001**
	groups*genre	0.19	-2937	1488	30.22	10	**<.001**
DC/T	groups	0.09	-44.89	27			
	genre	0.08	-190	103			
	groups+ genre	0.17	-227	123	41	2	**<.001**
	groups*genre	0.18	-238	139	30	10	**<.001**
CP/C	groups	0.008	-1074	542			
	genre	0.08	-1239	627			
	groups+ genre	0.09	-1238	629	3.98	2	0.13
	groups*genre	0.10	-1241	640	21.46	10	0.01
CP/T	groups	0.002	287	-138			
	genre	0.062	166	-75			
	groups+ genre	0.064	169	-74	0.95	2	0.62
	groups*genre	0.07	165	-62	23.72	10	0.008
CN/C	groups	0.008	1239	-614			
	genre	0.14	945.75	-464			
	groups+ genre	0.151	945.80	-462	3.95	2	0.13
	groups*genre	0.159	946.23	-453	19.56	10	0.03
CN/T	groups	0.02	2646	-1318			
	genre	0.09	2405	-1194			
	groups+ genre	0.12	2398	-1189	11.09	2	**0.003**
	groups*genre	0.13	2386	-1173	32.24	10	**<.001**
VP/T	groups	0.0548	1452	-721			
	genre	0.0542	1351	-667			
	groups+ genre	0.10	1332	-656	22.64	2	**<.001**
	groups*genre	0.12	1320	-640	32.22	10	**<.001**

– Degrees of freedom for the groups-only model is 5, for the genre-only model is 8, for the groups + genre model is 10, and for groups * genre model is 20.

The findings of four mixed-effects models and their comparisons for the syntactic measures also reveal very similar patterns with regard to the best fitting models which explain most of the variations in the values of the syntactic indices as illustrated in table 6.38. The model fit indices of AIC and R^2 also show the best fitting model is the model with the interaction of groups and genre as the fixed effects, and the next best-fitting model is the model with the additive effects of these two predictors. The difference between these two models is significant for most of the indices, i.e., C/T (clauses per T-units), CT/T (complex T-units per T-unit), DC/C (dependent clauses per clause), DC/T (dependent clauses per T-unit), CN/T (complex nominals per T-unit), and VP/T (verb phrases per T-unit) as indicated by the chi-square tests of model comparison when the Bonferroni-corrected alpha level of 0.004 is applied to the bootstrapped p-values. For the rest of the measures, this difference is only significant at the 0.05 level and not at the stricter 0.004 level. With regard to the interactional model, the largest R^2 value among the syntactic indices is recorded for both CT/T and DC/C (R^2 = 0.19) and the smallest R^2 value is recorded for the CP/T (coordinate phrases per T-unit, R^2 = 0.07). These results are also consistent with smaller AIC values for these indices.

The interaction of groups and rhetorical sections (the EFL group and abstract sections taken as the reference levels) resulted in an increase in the estimates of the measures of CT/T and DC/C from EFL abstracts to method sections of ESL group, CN/C and CN/T from EFL abstracts to literature review sections of ESL and NS, and VP/T from EFL abstracts to ESL and NS method sections and ESL result sections as shown in the estimates of the interaction terms.

The difference between the two models which isolate the effects of groups and genre is a straightforward and consistent one. Since these two models are non-nested, AIC values are better suited to compare them. Between the two, there are the genre-only models that receive much lower AIC for all syntactic measures. This indicates that, between the effects of groups and genre, genre variations are more predictive of variations in all syntactic complexity measures.

Overall, the findings show that the model with the interaction effects of groups and genre is the best-fitting model for explaining the variations in both lexical and syntactic indices, followed closely by the model that examines the additive effects of these two predictors. This important finding could be beneficial for future research to include both predictors in the studies of lexical and syntactic proficiency of postgraduate academic writing.

6.6.3. A Summary of Key Findings of 6.6

In this section, the effectiveness of including the ‘groups’ of students with different English language backgrounds and academic contexts as well as the rhetorical sections of dissertations as the main predictors of lexical and syntactic complexity of MA dissertations is examined. This is to investigate the extent to which the type of lexical and syntactic complexity measures (and by extension the amount of lexical and syntactic structures) can be attributed to a main text-extrinsic factor (groups of students) and a main text-intrinsic factor (rhetorical sections with different communicative purposes) in postgraduate academic writing.

The findings show that the models with the interaction effect of groups and genre/rhetorical sections explain most of the variations in the values of lexical and syntactic measures. This shows that the values of most of lexical and syntactic complexity measures for each groups of students heavily depend on the type of text, i.e., its rhetorical section. An instance that was discussed is the case of lexical sophistication in introduction sections where the two groups of English L1 and ESL produced more non-repetitious sophisticated words than the EFL group compared to other rhtorical sections. The interaction effects of groups and rhetorical sections are most significant in the abstract and literature review rhetorical sections for the lexical indices, and in the abstract and result sections for the syntactic indices as indicated by the fixed effects' estimates and their corresponding bootstrapped p-values. Between 'groups' and 'rhetorical sections', the models with the isolated effect of genre/rhetorical sections capture more variations in all lexical and syntactic complexity measures investigated in this study. This indicates that the variations in all lexical and syntactic complexity measures depend more on the rhtroical function/section of a text than the group of students who produced the texts. This shows the importance of this text-intrinsic feature and significant variations among different rhetorical sections regarding the type and amount of various lexical and syntactic structures.

These findings, overall, suggest that these two variables are inter-dependent when it comes to their predictability power that accounts for the variations in these complexity indices, and by extension the base production units and structures.

6.7. Random Forest Predictive Classification modelling: Finding the Strongest Predictors of Rhetorical-section and Group Memberships

Previous sections presented two statistical modelling methods to scrutinising the lexical and syntactic datasets. In 6.5 structural factor analyses were conducted which first examined the existing theories/categories regarding the structure of the lexical and syntactic indices via a series of structural equation models and then further explored the structure of this study's measures through exploratory factor analyses. Once a new picture of the overall and specific structures of the data was obtained, a series of linear mixed-effects models were then built in section 6.6 to find the effects of groups and rhetorical sections on the values of lexical and syntactic complexity measures and to determine which indices explain greater amounts of variances in the datasets considering the three groups and six rhetorical sections.

In this section, the final statistical modelling method of this study is presented which takes advantage of the available machine learning (ML) classification algorithms in R for prediction purposes. Analyses of variance have already established group differences for these measures, but does not allow making conclusions about which measures best predict proficiency and rhetorical sections. This section focuses on disentangling the relative contribution of the different measures in this regard. An important upshot of this section is to determine the most important predictors of lexical and syntactic proficiency among the 22 lexical and 11 syntactic complexity indices for groups of students with different English language backgrounds and rhetorical sections with various communicative purposes, using the variable importance features in various ML algorithms. This is to assess the measures' contributions to predicting the response variable, in this study's case the groups of students and the rhetorical sections of their dissertations and the impact of each feature/variable on the model's decision. In this section I also build models to predict group membership as well as membership to any of the six rhetorical sections, given a set of values for the lexical and syntactic measures investigated in this study (i.e., to calculate the probability that a given value belongs to a specific category/group). To accomplish the mentioned goals, the random forest method is chosen as a supervised machine learning approach to classification.

Random forest has already been used in several linguistics research such as Tagliamonte and Baayen (2012), Brown et al. (2014), and Baumann and Winter (2018). Many other researchers (e.g., Boulesteix et al., 2012; Gries, 2019; Kuhn, 2008; Probst, Wright, & Boulesteix, 2019; Ziegler & König, 2014) also point out to a number of advantages in using a random forest algorithm over other predictive methods. Random forest is robust to multicollinearity (i.e., correlated predictors) because of using the 'Feature Bagging' which is the use of a subset of features/variables at every stage. This is particularly helpful in this study as previous sections showed high correlations among many of these complexity measures. Furthermore, random forest does not have the assumption of the normality of the data because of the bootstrapping method in sampling, and hence robust to outliers. It also internally does the data standardisation which is important when the variables have different scales/metric. Random forest is widely used as a feature selection method, i.e., it selects the most important variables as predictors, which is useful specifically in studies with many variables and a relatively smaller number of observations. Random forests can solve both classification and regression problems in models called 'CART' models and can handle non-linear relationships. Gries (2019) particularly recommend this approach in corpus linguistics research as the naturally-occurring data (e.g., texts) "are often (extremely) Zipfian distributed" that might cause convergence problems in linear regression models, and because of the 'collinear' nature of predictors which might lead to "unstable regression coefficients" (p. 2). He maintains that random forest as a nonparametric alternative could overcome these problems of data sparsity and collinearity in corpus-based research.

Before using random forest and fitting models, a training dataset (usually 70% of the original data) and a testing dataset (or validation data that is the remaining 30% of the original data) need to be specified by the researcher.

A typical random forest algorithm follows these steps. The random forest algorithm takes the original training dataset (this 70% of the original data) and internally makes many alternate versions of the training set using sampling with replacement. This process is also called Bagging' or 'Bootstrap Aggregating for creating 'random samples with replacement' of the data. The sampled data is called the 'bagged sample' and any data not contained in this set (about 32% of the training set, on average) is called the 'out-of-bag' sample. These are analogous to simulated training/test set splits.

A random forest is made up of decision trees. A decision tree is a tree-like structure that follows an if-else rule to partition the data into distinct parts. Each decision tree takes its training data to learn which data point (e.g., a lexical or syntactic value) belongs to which class (e.g., different groups or rhetorical sections). In random forest, a random subset of variables (e.g., a random subset of lexical or syntactic measures) is evaluated each time a split is made. The algorithm then creates a separate decision tree for each of the bagged samples. This collection of many decision trees is called a forest. When predicting a data point to each class, each tree in the forest makes a prediction and a simple voting count is used to make the final prediction. That is, the final classification prediction is based on the averaging of all the trees' classification predictions.

To get a measure of how well a tree performs, each is used to predict its out-of-bag dataset (e.g., in this study's case how well it predicts the classes of groups and rhetorical sections) and a measure of performance is computed. The average of these performance statistics is an estimate of how well the forest would perform on a future/unseen datapoint. This overall average of prediction scores is then used on the original testing dataset (e.g., the 30% of the original data) to validate the accuracy of the prediction of random forest. For a detailed conceptual and practical introduction to random forest see Kuhn (2008) and Probst et al. (2019).

In this study, the two R packages of ranger (version 0.11.2, Wright & Ziegler, 2017) and caret (version 6.0-84, Kuhn, 2008) were used to build four random forest models on the lexical and syntactic datasets to predict rhetorical-section and group membership and to find the most important lexical and syntactic predictors. This is because algorithms in different packages have nunaced differences in computation methods that can be helpful for a specific type of data. This practice is not discouraged so long as one records the parameters for training the data as well as performance indices on the validation process, as will be explained more. Four models were built and the best-performing one was chosen: three using the caret package and one using the ranger. For each model, the model performance indices of accuracy, precision, recall (sensitivity), specificity (true negative rate) and F1 are recorded from the confusion matrices. To avoid extra technical details here, an explanation of the quantification methods of these performance indices is presented in Appendix C2. It is noteworthy that a model's parameters directly affect these performance indices and the variables selected as important predictors by the model. Probst et al. (2019) elaborate on this issue and recommend the practice of hyperparameter tuning which is optimising each model's parameters separately to achieve higher accuracy and to avoid overfitting. This parametrization process is carried out in this study based on the recommendations of Probst et al. (2019) for each model by changing the default settings of the functions and recording the performance indices to select the best performing set of parameter values associated with higher model accuracy. The parameters which were tuned in this study are the number of trees, minimum node size, sampling parameters, variable importance measure, and the splitting rule.

The models take the groups and rhetorical sections as the response variables for classification purposes and 22 lexical and 11 syntactic measures as predictors; hence four separate models. To comply with the assumption of independence of data points, the aggregated datasets (i.e., six rhetorical sections aggregated as weighted mean as explained in 6.1) are used to construct the lexical and syntactic models to predict group membership. The results of these models and the graphs of most important predictors in each model are presented in 6.7.1 and 6.7.2 for the lexical and syntactic datasets respectively.

6.7.1. Lexical and Syntactic Predictors of Group Membership

Unlike the structural equation modelling and mixed-effects modelling in previous sections, in classification models, the complexity measures were taken as predictors. Therefore, to find the strongest lexical and syntactic predictors of group membership, two models were built using random forest multi-class classifier (as I explained in detail in 6.7) on both lexical and syntactic datasets.

The first model seeks to predict group membership given the values of 22 lexical measures investigated in this study. In other words, the algorithm takes each value/observation of each lexical measure in the test data and predicts whether that value belongs to the EFL, ESL, or the English L1 group based on the trained model. The second model is similar to the first model (i.e., prediction of group membership) but for the 11 syntactic measures. Table 6.39 reports the performance indices of these two predictive models and graphs 6.11 and 6.12 demonstrate the most important lexical and syntactic measures as predictors of groups of students respectively.The most accurate models based on the performance indices were achieved when training parameters were specified according to the information in the caption of table 6.39. The vertical axis in the left side of each graph represents the overall important variables. The specific predictors of groups of students based on these complexity measures can also be found in segregated columns labelled as 'EFL', 'ESL', and 'English L1' to answer the relevant research questions in 6.8.

Table 6.39. Performance Indices of the most-accurate random forest models for predicting group membership

Model	**Accuracy [CI]**	**Precision**		**Recall (Sensitivity)**	**Specificity (True Negative Rate)**	**F1**
Groups ~ 22 Lexical Measures	54% [40, 67]%	EFL	69%	52%	88%	59%
		ESL	38%	40%	70%	39%
		NS	58%	68%	73%	62%
Groups ~ 11 Syntactic Measures	51% [38, 64]%	EFL	51%	81%	62%	63%
		ESL	38%	25%	81%	30%
		NS	59%	45%	83%	51%

– Parameter setting for the lexical dataset: 10-fold repeated cross-validation with 3 repeats, 147 samples with an average sample size of 132, and the number of variables for splitting at each node set to 22 on the genre-aggregated lexical data in the caret package.
– Parameter setting for the syntactic dataset: 3000 trees and 3 variables for splitting at each tree node, the 'extratrees' method as the splitting rule, 3 minimum node size, 147 samples with an average sample size of 63 on the genre-aggregated syntactic data in the ranger package.

The model that predicts group membership based on the values of 22 lexical measures received its best performance with 54% accuracy and a CI of [40, 67]% accuracy. This indicates that if we were to collect another sample from the same population, this model could still predict the group of students based on the given lexical values with a maximum of 67% accuracy. This is well above the chance level of 33% if the data were to be classified purely randomly. Among the three groups, however, the English L1 group was more accurately predicted (i.e., less overall type I and II error rates associated with assigning the lexical values to the English L1 group). It also shows that the random forest algorithm had a hard time correctly classifying the ESL group based on the given values in the test dataset (F1 score = 39%). This could be due to similar performances of ESL and English L1 groups in producing lexically dense, diverse, and sophisticated texts as demonstrated by the results of the linear models.

To test this hypothesis, I ran another random forest test keeping the main parameters the same but omitting the ESL group from the 'groups' factor. The accuracy of this model jumped from the 54% of the full model to 78% with a CI of [63, 90]% accuracy of the second model which clearly shows that the model is fundamentally a very good one and the low accuracy of the full model with the three groups could be due to the similar performances of the ESL and English L1 groups as indicated by ANOVA tests in 6.3. To disambiguate this, i.e., to examine whether the higher accuracy of the second model was because of the fewer number of classes (2 classes) or due to data similarity between the ESL and English L1 texts, I also ran the third model with 2 classes, but with the similar performing groups of ESL and English L1. This third model's accuracy was recorded as 50% with a CI of [34, 66]% which is much lower than the second model, confirming my hypothesis that the original model's lower accuracy for the ESL texts could be primarily because of data similarity with the English L1's. The full results of these second and third models will be presented in Appendix B.

The top three overall predictors of correct group classification of the full model as illustrated in graph 6.11 (the vertical axis of variable importance, see Appendix C2 for more details) are recorded for the mattr (moving-average type-token ratio), ld (lexical density) and msttr (mean segmental type-token ratio) measures. As for the individual effects of the lexical measures on the prediction accuracy of group classification, the top predictor of the EFL and ESL groups is the mattr index and the top predictor of the English L1 group is the ndwesz (number of different words, first type) index followed closely by mattr.

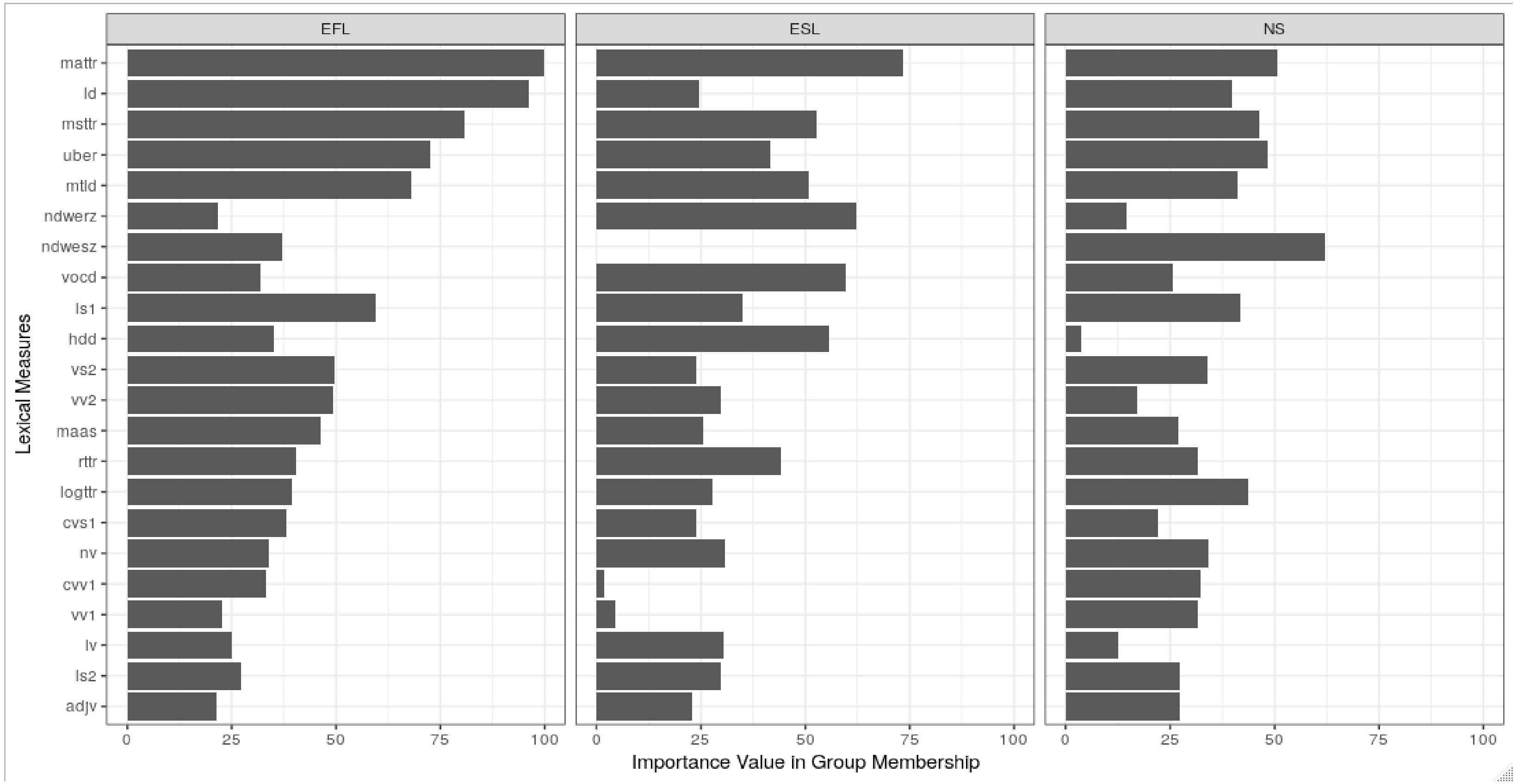
In the full syntactic model (with the three groups), the top three overall important syntactic measures in predicting group membership are CT/T (complex T-units per T-unit), DC/C (dependent clauses per clause), and CP/C (coordinate phrases per clause). This is somewhat consistent with the results of the mixed effect models for the effect of groups on the values of syntactic indices. Table 6.38 indicates that the CT/T and DC/C measures both explain larger amounts of variations compared to other indices as demonstrated by the R-squared values. However, the CP/C index is only a distinguisher of group performance based on the ANOVA results, where the EFL group outperformed the other two groups with a medium effect size. The CP/C index, nevertheless, received a large factor loading as a distinct factor (along with CP/T or coordinate phrases per T-unit) in the results of the exploratory factor analysis (table 6.34). These could be possible reasons for this measure to be selected as one of the most important predictors of group membership by the random forest algorithm. These findings with regard to the CP/C (coordinate phrases per clause) index reveals how the three statistical tests of ANOVA, factor analysis, and random forests show different aspects of the measures' patterns in datasets and together help to explain the role a particular index performs compared to other indices.

This model that predicts the group classification based on the values of 11 syntactic measures reached its highest accuracy of 51% and a CI of [38, 64]% which shows that these indices can moderately predict which group a given student's value belongs to. A close look at the values of other model performance indices of precision, recall and F1 also shows a similar overall accuracy trend.

The random forest algorithm in the ranger package could more precisely predict the English L1 group (less type I error for the English L1 group) but the overall correct identification of both true positives and false negatives as expressed via recall is higher for the EFL group. Consequently, the EFL group receives the largest value of the F1 score (63%). Once more, the algorithm could less accurately predict which values belong to the ESL group that could be due to similar performances of the ESL and English L1 groups as mentioned before.

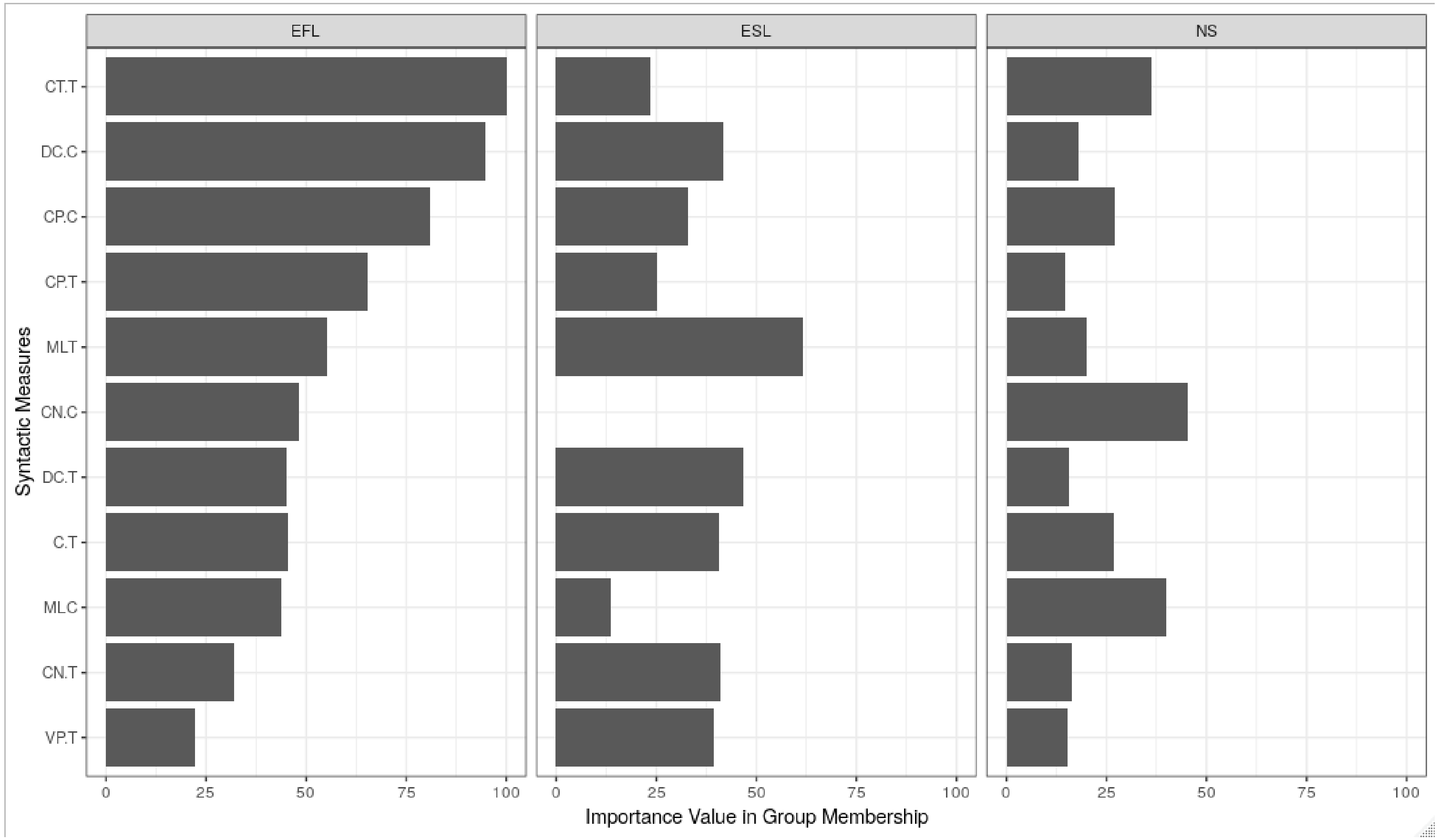
To test this hypothesis, I ran a second random forest with the main parameters of the full model but omitting the ESL group from the 'groups' factor. The accuracy of this secondary model jumped to 57% with a CI of [41,72]%. Once more, this significant increase in the model accuracy indicates that the lower accuracy of the full model with the three groups was due to the similar performances of the ESL and English L1 groups as demonstrated in section 6.3 with the ANOVA tests. To disambiguate this situation as well, I ran a third model with the 2 classes of ESL and English L1 while keeping other parameters the same. This third model obtained an accuracy of 47% with a CI of [32, 63]% which are much below the second model, once more confirming the hypothesis that the number of classes has not caused higher accuracy of the second model, but that similarity vs dissimilarity of the data between the groups being compared is the main reason for accuracy values. The full results of the second and third models will also be presented in Appendix B.

Graph 6.11. Important lexical predictors of group membership



– The vertical axis represents the overall strong lexical predictors of group membership for all groups. The horizontal axis represents the variable importance values internally computed by the random forest algorithm (more details in Appendix C).

Graph 6.12. Important syntactic predictors of group membership



– The vertical axis represents the overall strong syntactic predictors of group membership for all groups. The horizontal axis represents the variable importance values internally computed by the random forest algorithm (more details in Appendix C).

These two scenarios for the differences of accuracies of 3-class vs. 2-class models for both lexical and syntactic datasets for classifying group memberships can happen when all/most values of a group (e.g., ESL in this study) is similar to a subset of another group (e.g., English L1); therefore, the different part in the English L1 data gets easily classified by the algorithm which results in overall higher accuracy for the English L1 group. However, the similar part between the two groups do not get correctly classified for either of the groups, but it affects the ESL group more, and this results in lower accuracy for the ESL class. This can also be reflected in how the algorithm selects a certain measure as a top predictor. As mentioned earlier, the classification for each group (in separate panels) is based on how accurately the algorithm can assign the data points of a particular measure to the group it belongs. Due to the quantification method of a particular measure and different or similar performances of the groups, it is possible that a measure can accurately classify a group but not the other ones. This can be seen, for instance, for the ndwesz measure which is based on 10 random samples of 50 consecutive words for which the students of a group as a whole may not have produced varied words within the 50-word random samples compared to the other groups (e.g., as opposed to mattr where the lexical diversity of the whole text is considered not just sub-samples). In the case of ndwesz, the ESL and Eglish L1 groups performed very similar, but the English L1s have slightly larger values; as a result, the algorithm could have disproportionately classified the similar data points in favour of the English L1s. However, the graph for the few top predictors (mattr, ld, msttr, uber) shows a more proportionate classification in the case of lexical measures because of more differences in the performances of the three groups.

The results of this section showed that lexical density and the two lexical diversity measures based on word strings (mattr and msttr), alongside coordinate phrases per clause (CP/C) and the two subordination indices that mainly gauge dependent clauses can better distinguish the writings of the three groups and, therefore, are chosen as better predictors of lexical and syntactic complexity, and by extension proficiency differences of students with different English language backgrounds.

6.7.2. Lexical and Syntactic Predictors of Membership to Rhetorical Sections

The same procedures and principles that were outlined in 6.7 and 6.7.1 were also followed for predictive classification models to find strong lexical and syntactic predictors of membership to each of the six rhetorical sections of dissertations. The lexical model seeks to predict the classification of the six rhetorical sections of MA dissertations given the values of 22 lexical indices and the syntactic model predicts the membership to any of the six rhetorical sections specified in this study. Table 6.40 presents the performance indices for these two models as derived from the confusion matrices of the respective models. The most accurate models were achieved when the parameters for training the data were tuned according to the information in the table caption.

The model that specified the lexical indices as predictors of rhetorical sections received a high accuracy of 59% with a CI of [54,64]%. This finding suggests that the lexical indices investigated in this study are relatively good predictors of rhetorical sections of master's dissertations of the three groups. The top three overall important predictors in this model are rttr (square root of type-token ratio), logttr (logarithm of type-token ratio), and ls2 (lexical sophistication type two). This finding is highly consistent with the results of mixed-effects models presented in table 6.37 where these three variables were shown to explain the largest amounts of variation (R-squared) in the models that examined the effect of the rhetorical section alone on the values of lexical indices.

Table 6.40. Performance Indices of the most-accurate random forest models for predicting membership/classification of rhetorical sections

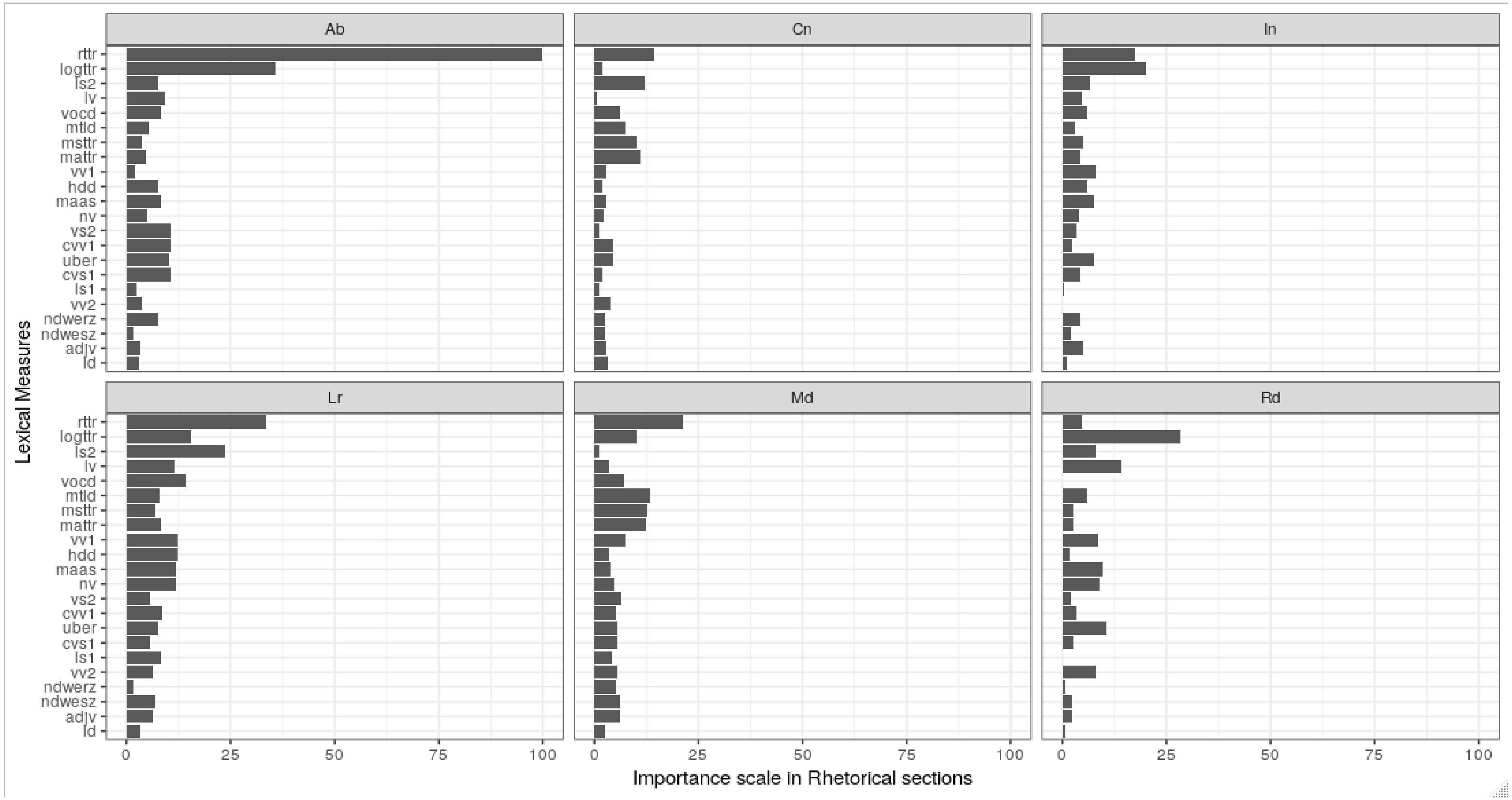
Model	**Accuracy [CI]**	**Precision**		**Recall (Sensitivity)**	**Specificity (True Negative Rate)**	**F1**
Genre ~ 22 Lexical Measures	59% [54, 64]%	Ab	93%	95%	99%	94%
		In	52%	57%	88%	55%
		Lr	61%	67%	92%	64%
		Md	44%	53%	87%	48%
		Rd	55%	48%	92%	52%
		Cn	53%	39%	92%	45%
Genre ~ 11 Syntactic Measures	35% [30, 40]%	Ab	56%	51%	92%	53%
		In	25%	27%	86%	26%
		Lr	41%	34%	90%	38%
		Md	45%	39%	90%	42%
		Rd	26%	40%	78%	31%
		Cn	21%	17%	85%	19%

– Rhetorical sections are labelled as 'genre'.
– Parameter setting for the lexical dataset: 5-fold cross-validation, 12 variables for splitting at each tree node, and 882 samples with an average sample size of 706 in the caret package.
– Parameter setting for the syntactic dataset: 5-fold cross-validation with 3 repeats, 882 samples with an average sample size of 706, and 6 variables for splitting at each node in the caret package.

This result is specified across classes (i.e., all groups together) for each rhetorical section as presented in graph 6.13. Regarding the top predictors in each rhetorical section, the rttr index better predicts the abstract, literature review, method, and conclusion classes and the logttr measure better predicts the introduction and result sections. As we can notice, a measure like rttr is a good predictor of diversity of lexis in both short abstract sections as well as long literature review sections, indicating that text length did not have a significant impact on the type of predictors. Due to multiple classes (i.e., the six rhetorical sections) and the varying text length among them, it is imperative that we take into account other performance indices besides accuracy. The F1 score is another important classification performance index which is a harmonic mean of recall and precision; in other words, it takes into account ‘all’ as well as ‘only’ correct classifications in each class/rhetorical section. This index shows that the classifier predicts the abstract section with a staggering 94% accuracy, while the other sections received between 45 to 64%. This is an important point which tells us that the three measures of rttr, logttr, and ls2 correctly and better classify the abstract section among all six rhetorical sections across the three groups. By the same token, the lowest value of F1 score that is recorded for the conclusion section (45%) means that the algorithm could less accurately predict whether a given lexical value belongs to the conclusion section.

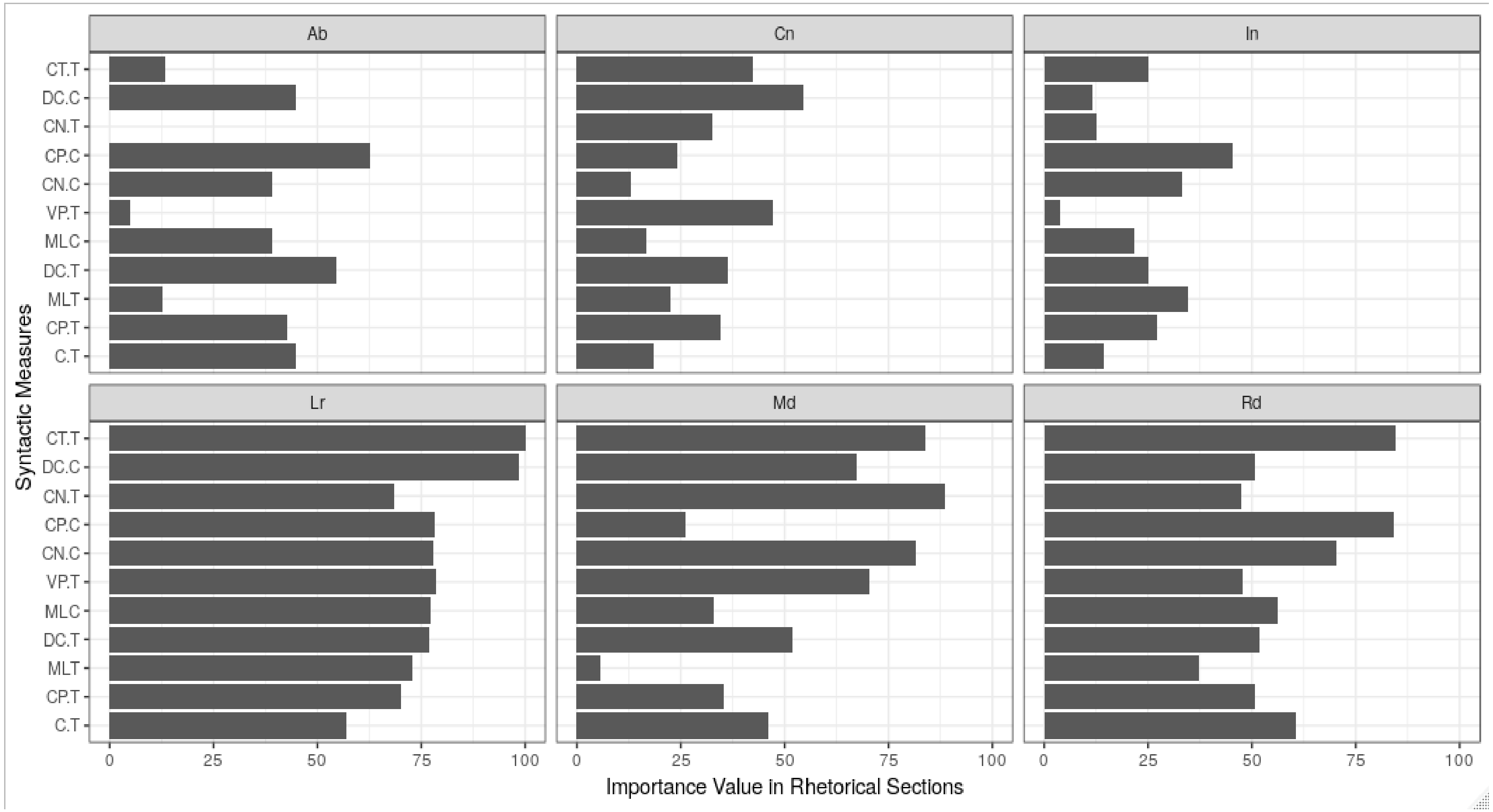
The last random forest model is built to classify the six rhetorical sections based on the values of the syntactic measures. This model received its peak overall performance with 35% accuracy and a CI of [30,40]%. Table 6.40 shows similar values for other performance indices. Considering that this obtained model accuracy is well above the chance level of 17%, these findings indicate that the syntactic measures investigated in this study are reasonably good predictors for correctly classifying the rhetorical sections of master’s dissertations (e.g., the investigated syntactic structures are a function of/depend on the rhetorical aspects of academic texts). In other words, the algorithm could reasonably distinguish different values of syntactic measures across the three groups based on their actual respective categories of rhetorical sections. The abstract section again receives a better prediction accuracy compared to other rhetorical sections based on the values of all performance indices. Regarding the overall top predictors as demonstrated in graph 6.14, the two measures of CT/T (complex T-units per T-unit) and DC/C (dependent clauses per clause) are selected similar to the other syntactic-based model. This finding indicates that these two measures consistently show to be better predictors of group and rhetorical-section membership compared to the values of other syntactic indices in this study and hence recommended to future researchers.

Graph 6.13. Important lexical predictors of rhetorical sections



– The vertical axis represents the overall strong lexical predictors of rhetorical-section membership. The horizontal axis represents the variable importance values internally computed by the random forest algorithm (more details in Appendix C).

Graph 6.14. Important syntactic predictors of rhetorical sections



– The vertical axis represents the overall strong syntactic predictors of rhetorical-section membership. The horizontal axis represents the variable importance values internally computed by the random forest algorithm (more details in Appendix C).

In these predictive models, the use of bootstrap (re)sampling and confidence interval methods which mimic the population of the study outside the study sample at hand as well as the cross-validation method that is designed to assess the predictive power/accuracy of the model to a new dataset, indicate that the selected top lexical and syntactic indices would be relatively good predictors of rhetorical section and group memberships in a given corpus of postgraduate academic writing. In the absence of any other study with such predictive models, it is recommended that future researchers employ these lexical and syntactic variables selected by the algorithm in such predictive classification studies.

6.7.3. A Summary of Key Findings of 6.7

Random forest classifiers were used in this section to investigate which of the 22 lexical and 11 syntactic complexity measures are good predictors of the membership to groups and rhetorical sections. That is, how accurately the predictive algorithm could assign the values of these complexity measures to the groups of students and any of the six rhetorical sections that they belong. The findings show that all four specified models could classify these complexity measures' datapoints reasonably well. The accuracies of these models range from 54% and 51% for lexical and syntactic models for classifying the groups and 59% and 35% for lexical and syntactic models for classifying the rhetorical sections respectively. Being well above the chance level, these model accuracy values thus indicate that these lexical and syntactic complexity measures do relate to groups and sub-sections' complexity differences in a meaningful way. Taking these complexity measures as a proxy to proficiency, these results also suggest the proficiency differences of these groups and the lexical and syntactic features that can predict this proficiency both in the entire dissertations and in each of the rhetorical sections as sub-genres of these dissertations. Being below very high accuracy levels (e.g., below 90%), these model accuracy values also indicate that there are other linguistic features that could be incorporated into such models to improve the accuracies and to obtain a better picture of other possible predictors of complexity and proficiency differences.

Some of the lexical and syntactic complexity measures were also shown to be stronger predictors in the models. The rttr, logttr, and ls2 measure, for instance,s are shown to be stronger lexical predictors of membership to the rhetorical sections, especially the abstract section which obtained the highest classification accuracy among other rhetorical sections. This indicates that the values of these two measures of lexical diversity along with lexical sophistication index type II were distinctly high in some sections. Interestingly, rttr values could better predict/classify the short abstract sections as well as very long literature review sections suggesting more variations in the values of this index across rhetorical sections as produced by all groups. This result is consistent with the R-squared values for these measures in genre-only mixed-effect models. The three groups altogether also produced higher amounts of lexical sophistication in literature review and conclusion sections. The mattr, ld, and msttr indices were shown to be the best predictors of students' groups in all rhetorical sections combined. This finding complements the results of analyses of variance in that the EFL group produced the least lexically dense and diverse texts compared to the other two groups overall, and therefore, the random forest algorithm could much easily assign these distinctly lower values to the EFL group (higher classification accuracy as shown in table 6.36).

The syntactic-based models were less accurate compared to their counterpart lexical models. In both syntactic models, the CT/T and DC/C indices were ranked as the top syntactic predictors of group and rhetorical section memberships, suggesting that dependent clauses as indices of subordination are more effective in distinguishing between the production/performances of postgraduate students with different English language backgrounds as well as different rhetorical sections/aspects of postgraduate academic writing. The CP/C index was the next best index in classifying the syntactic values to their respective groups. This result is also reflected in the analyses of variance in that, overall the EFL group produced more coordinated phrases and therefore, the higher values could be more easily classified for this group as shown in graph 6.12. The CN/T index is also ranked as the third predictor of rhetorical sections of these students with better prediction accuracy across literature review and method & design sections.

Overall, the findings of this section complement those of analyses of variance and mixed-effects models but with the added benefit of predictability and taking these complexity measures as independent variables whose higher or lower values can predict a particular group or rhetorical section.

6.8. Addressing the Research Questions in Light of the Combined Results

In chapter one, I formulated four main groups of research questions, each with a few sub-questions. In this section, I will answer each question based on the findings of one or several of the analyses presented in sections 6.1 to 6.7. Since detailed discussions of the findings of this study and the comparisons with previous research have been presented in sub-sections of this chapter, in this section I only provide concise answers to the specific research questions while referring to the corresponding sections of the results with detailed discussions. The conclusions and implications of these findings will be further discussed in the next chapter in detail.

6.8.1. Answering Group A of Research Questions. Measure-selection process: Examining 22 Lexical and 11 Syntactic Complexity Measures

The first group of research questions deals with the measure-testing process to test the efficacy of each of the lexical and syntactic measures in capturing differences of academic texts, to find the relationship between them, to verify and explore the structures in these measures, and finally to determine the overall indicators and predictors of linguistic proficiency and performance to assist the measure-selection processes of future studies. The specific questions are formulated as:

A1. *How do the selected lexical and syntactic complexity measures compare with and relate to each other as indices of quality of academic texts at the postgraduate level in the whole corpus of this study? Is the construct-distinctiveness of these lexical and syntactic categories (see details in 6.4) confirmed with this corpus of MA dissertations (see details in 5.2)?*

The overall construct distinctiveness of the three lexical constructs of density, diversity, and sophistication are supported with this study's corpus of MA dissertations on the aggregated dataset (the whole corpus) as indicated by overall weak correlations between the measures belonging to these different constructs in table 6.21. Strong correlations, however, were noticed between the measures in the same construct as demonstrated by tables 6.21 and 6.22. The same pattern is observed in the syntactic dataset where the overall construct-distinctiveness of the subordination, coordination, and phrasal sophistication is indicated by the overall weak correlations in table 6.23 and strong correlations are observed between the indices in the same category/construct (tables 6.23 and 6.24). The construct-distinctiveness of these lexical and syntactic categories are also reported in Lu's (2012) study of oral narratives, Šišková's (2012) study of university students' written narratives, as well as the assumptions in Lu (2010) and Lu and Ai's (2015) studies. These findings imply that the construct-distinctiveness of these categories are independent of the mode of language and genre variations.

Regarding the three lexical constructs, the indices in lexical diversity have higher correlations with each other; this is more noticeable between various log-based and word-string based measures. The lexical diversity of TTR of word classes has weaker correlations with the other two-mentioned sub-constructs. In the syntactic dataset, a lower-than-expected correlation of r = 0.6 is observed between MLC (mean length of clauses) and MLT (mean length of T-units) measures which belong to the same syntactic construct of mean length of production units while strong correlations (r = > 0.8) were found between other syntactic indices belonging to their respective constructs. A detailed discussion of the relevant findings was provided in section 6.4.

The exploratory factor analysis on the aggregated datasets (table 6.30), however, show a few misplaced indices, for instance, the loading of ls2 (lexical sophistication type II) on the 4th factor beside logarithm-based measures of lexical diversity as discussed in 6.5.2 and the loading of ls1 (lexical sophistication type I) on the 7th factor beside lexical density that, as discussed, could be due to similar quantification methods and the fact that in the entire academic writing corpus, the students produced similar amounts of lexical words and sophisticated lexical words overall. Similarly, the two indices of MLT and VP/T did not load on the expected factors: both MLT and VP/T loaded with other indices that have T-units as their denominators in the 2nd factor (mainly subordination indices). Detailed discussions of these results and previous studies are presented in 6.5.4. In conclusion, the boundaries of these constructs seem not to be as rigid as the existing classifications in the literature suggest and that the mode (written vs spoken) and type (general SLA vs specialised academic writing) corpus have great impacts on the structure of these complexity measures and constructs.

A2. *To what extent do the selected lexical and syntactic complexity indices in this study fall into the current categories of lexical and syntactic constructs proposed in the literature (see 5.3.1)? What new structures are detected regarding this study's corpus of academic texts (the results of exploratory factor analyses)?*

Following the mantra “first confirm, then explore” recommended by McArdle (2011, p. 335) and after fulfilling the assumptions of factor analysis in both lexical and syntactic datasets, several confirmatory factor analyses using structural equation modelling were conducted to examine the current lexical and syntactic models in the literature (for details refer to sections 6.5.1 and 6.5.3). Two main lexical models (one with an overall classification and one with a fine-grained classification) along with their sub-models (to test and rule out the effect of multicollinearity of observed variables) were constructed. None of these models produced acceptable fit indices as demonstrated in tables 6.29 and 6.33 which suggests that these models' structures are not completely consistent with the proposed models in the literature.

An exploratory factor analysis was then conducted to find the nuances of the patterns/structures of the lexical indices. The seven-factor output model shows overall compliance with the current models in the literature but four lexical indices did not line up on the expected factors, suggesting that the structure of lexical measures representing various constructs are to some extent different in a postgraduate academic writing corpus (e.g., this study) vs. the ones in SLA and lower-level academic corpora (e.g., the existing classifications in Lu, 2012 and Ai & Lu, 2015). For detailed discussions of these findings see section 6.5.2.

Following the same statistical procedures for the syntactic dataset, one main syntactic model with two sub-models were specified and confirmatory factor analyses were conducted. None of these models produced acceptable values of fit indices either (see section 6.5.3) which could be due to very high correlation values between all syntactic indices in this study overall. This necessitated the further exploration of the dataset with exploratory factor analysis. The four-factor model specified in this stage shows an overall consistency with the model proposed by Lu and Ai (2015) concerning the syntactic constructs and their respective measures, except that the two measures of MLT (mean length of T-units) and VP/T (verb phrases per T-unit) lined up in an unexpected factor. A detailed discussion of these findings can be found in section 6.5.4.

A3. *Which lexical and syntactic constructs and measures can better capture differences in academic texts produced by three groups of postgraduate students (see details in 5.2.1 and 5.3.1) and what are the overall lexical and syntactic indicators of linguistic proficiency and performance as specified by between-group differences (see details in 6.3)?*

The three lexical measures of mattr (moving average type-token ratio), msttr (mean segmental type-token ratio), and mtld (measure of textual lexical diversity) which are all lexical diversity measures based on word strings as well as the five syntactic measures of MLT (mean length of T-units), C/T (clauses per T-unit), CT/T (complex T-units per T-unit), DC/C (dependent clauses per clause), and DC/T (dependent clauses per T-unit) which mainly target the subordination structures, consistently captured between-group differences across rhetorical sections with varying text length, are shown to be good indicators/discriminators of linguistic complexity (and by proxy proficiency) in postgraduate academic texts produced by English L1 vs L2 writers. However, care needs to be taken regarding some overall measures, such as CT/T, as Kyle (2016) mentioned, and other fine-grained indices may also need to be used in conjunction with these global measures to better capture proficiency differences in such studies. Since in this study the English L1 group was shown to have higher overall lexical and syntactic proficiency in academic writing, the indices with larger values for the English L1 group and consistently lower values for both EFL and ESL groups, could be considered as good indicators of linguistic proficiency (lexical and syntactic proficiency) and performance overall, and specifically in assessing proficiency-level and performance differences. These indices are indicated in tables 6.5 to 6.18.

A4. *What are the overall lexical and syntactic predictors of linguistic proficiency and performance of the groups as obtained from the predictive models (see details in 6.7)?*

The three lexical complexity indices of mattr, ld, and msttr followed closely by uber and mtld measures were selected by the random forest algorithm in the *caret* package to be the top predictors of lexical proficiency across groups and the top predictors of group membership by the random forest predictive model (i.e., how well these indices could assign each value to the group it actually belongs). This result is demonstrated in graph 6.11 where the indices in the vertical axis show the order of maximum importance of variables across classes (i.e., students' groups). Three of these measures, mattr, msttr, and mtld, as explained in the answer to research question A3, also better captured between-group differences. Most of these measures are indices of lexical diversity, especially the sub-construct of lexical diversity based on word strings; this indicates that lexical diversity (among the three constructs of density, diversity, and sophistication) is a better predictor of proficiency-level and performance differences in postgraduate academic writing produced by students with different English language backgrounds.

The top predictors of syntactic proficiency and the top syntactic predictors of group membership by the random forest predictive model are CT/T (complex T-units per T-unit), DC/C (dependent clauses per clause), and CP/C (coordinate phrases per clause) followed closely by CP/T (coordinate phrases per T-unit) and MLT (mean length of T-units) indices. Three of these measures (i.e., CT/T, DC/C, and MLT) were also found to capture between-group differences better (in the answer to research question A3). This finding suggests that the amount of subordination as indicated by the two subordination indices of CT/T and DC/C plays a major role in predicting the proficiency-level and performance differences in postgraduate academic writing of students with different English language background (e.g, English L1 vs. L2): higher-proficiency students (e.g., English L1 group) produced more subordinate structures and the lower-proficiency students (i.e., the EFL group) produced more coordinate structures (see for example the discussion of previous works presented in 6.7.1 and the discussions in Chen, Alexopoulou, & Tsimpli, 2019).

A5. *Which of the mixed-effect models explain the largest amounts of variation in the lexical and syntactic complexity indices in the whole corpus (see details in 6.6)?*

Among the four linear mixed-effects models specified in 6.6.1 and 6.6.2, the model with the interaction effects of groups and rhetorical sections (labeled as ‘genre’ in the tables) explained the largest amounts of variations for most lexical and syntactic indices investigated in this study as indicated by smaller AIC and larger R-squared values. The R-squared among other fit indices reported in tables 6.37 and 6.38 is an intuitive measure of the overall fitness of the model and the strength of the relationship between the model and the response variable (e.g., any of the lexical and syntactic indices). For this specific type of model, the highest R-squared value is found for the rttr measure of lexical diversity ($R^2 = 0.59$; e.g., nearly 60% of the data for the rttr index is explained around its mean) and the lowest value is recorded for the adjv index (adjective variation from the sub-construct of lexical variation based on the type-token ratio of word classes) in the lexical dataset. The rttr index (as indicated by table 6.12 and discussed in the answer to research question C1) is also the top overall lexical predictor of different rhetorical sections (i.e., the membership to any of the six rhetorical sections based on the values of 22 lexical indices) and the adjv index is among the last two predictors (very low prediction power).

The largest R-squared value in the syntactic dataset is found for the CT/T (complex T-units per T-unit) and DC/T (dependent clauses per T-unit) indices with $R^2 = 0.19$ (e.g., nearly 20% of the variability in the CT/T and DC/T measures is concentrated around their means) and the smallest value is recorded for the CP/T (coordinate phrases per T-unit) measure with $R^2 = 0.07$. The CT/T index is also the top predictor of both rhetorical section and group membership as indicated by graphs 6.13 and 6.14 and the CP/T is among the last two predictors of membership to rhetorical sections (very low predicting power to correctly classify the values into the rhetorical sections they belong to).

These observations across the findings of different statistical analyses suggest that there is a strong relationship between the top or low predictors of lexical and syntactic proficiency in non-aggregated datasets (i.e., data separated by rhetorical sections and/or groups) and the amounts of variations explained by the interactional model in the aggregated/ entire dataset. The findings of mixed-effects modelling also indicate that the two variables of groups of students with different English language backgrounds and rhetorical sections with different communicative purposes are inter-dependent when it comes to the predictability power regarding the variations in most lexical and syntactic complexity indices.

6.8.2. Answering Group B of Research Questions: Differences of English Academic Texts Written by English L1 vs L2 (EFL and ESL) Postgraduate Students

This group of research questions deals with the comparison of academic writing performance and proficiency differences of the three groups of EFL, ESL, and English L1 postgraduate students (e.g., to revisit the assumptions of the differences of English academic texts from L1 vs. L2 writers). The pedagogical implications of the findings will be thoroughly discussed in the final chapter. This group contains the following sub-questions:

B1. *Which group of students produced the most linguistically-complex texts, e.g., more lexically and syntactically complex texts (i.e., with larger values of each and/or all of the lexical and syntactic complexity measures and constructs selected in 5.3.1)?*

Overall, the English L1 group followed closely by the ESL group, produced more lexically and syntactically complex texts as specified by different lexical and syntactic constructs and measures in this study, and hence the most linguistically-proficient group. The EFL group produced the least linguistically complex texts overall regarding the production of lexically and syntactically complex structures. These results are consistent with the findings of the twenty-one studies reviewed in Ortega (2000). Taking complexity as a rough proxy to proficiency, the combined results of this study show that the English L1 group's text characteristics are closer to the trajectories of lexical and syntactic complexification of proficienct writers as discussed in detail in chapter three.

With respect to lexically complex texts, this distinction is more prominent in abstract, introduction, results, and conclusion rhetorical sections where the English L1 and ESL groups produced more lexically dense and diverse texts, while the EFL group outperformed these two groups in the production of lexically diverse texts only in the literature review sections with medium low-effects.

The English L1 and ESL groups also produced larger amounts of subordination structures which are believed to be indicators of syntactic proficiency; on the other hand, the EFL group produced more coordination structures that are believed to be indicators of lower syntactic proficiency learners. These findings are consistent with the findings of Grant and

Ginther (2000), Mancilla et al., (2015), Monroe (1975), Bardovi-Harlig and Bofman (1989) and Chen, Alexopoulou, and Tsimpli (2019) among others. For detailed discussions of the relationship between proficiency and linguistic complexity see section 1.3; for detailed findings see section 6.3.

B2. *To what extent do the EFL and English L1 students/groups differ regarding the production of lexically and syntactically complex texts overall and specifically (e.g., based on the six rhetorical sections)? Do any such differences have implications for EFL academic writing practices?*

The results of this study point to an overall pattern in lexical and syntactic differences of the English L1 and EFL groups as well as the specific differences in each rhetorical section. This latter case is an instance of the (un)awareness of form-function relationships of linguistically complex structures and their rhetorical functions in academic writing (e.g., in the discussion s of Lu et al., 2020) that has been explained in chapter three and will further be discussed in chapter seven.

Most of the lexical complexity measures which showed between-group differences as indicated by tables 6.7 to 6.13 marked the English L1 students as a more proficient group compared to the EFL group regarding the production of more lexically and syntactically complex texts. These distinctions are more noticeable regarding the mattr, msttr, and mtld indices which belong to the sub-construct of lexical diversity of word strings as well as ld (lexical density) with medium to large effect sizes. However, the EFL group outperformed the English L1 group regarding the values of lv (lexical variation), nv (noun variation), and vv2 (verb variation type II) indices with lexical tokens as the denominator with low-medium effects mainly in the literature review section. The EFL group also produced more lexically sophisticated texts as measured by Ls1 index with medium effects in the method, results, and conclusion sections while the English L1 students produced more lexically sophisticated texts (calculated via Ls2) only in the introduction section.

Overall, the findings suggest that the English L1 students produced more lexically dense and diverse texts than the EFL group and the EFL group produced more lexically sophisticated texts based on sophisticated words outside the frequently-used words (see details in 3.5.1.1) in three mentioned sections. However, as discussed earlier, a detailed qualitative analysis of these texts are needed to examine if these sophisticated lexical items are purely field-specific terminology.

With regard to the syntactic structures, the English L1 students consistently produced greater amounts of subordination structures such as C/T (clauses per T-unit), CT/T (complex T-units per T-unit), DC/C (dependent clauses per Clause) and DC/T (dependent clauses per T-unit) which are mainly clauses and dependent clauses with medium to large effects, as well as the length of T-units (MLT) with small to medium effects than the EFL group. The indices in the construct of phrasal sophistication marked mixed results for these two groups: the CN/T (complex nominals per T-unit) index recording larger values for the English L1 group with small to medium effects in the abstracts and literature review sections while the CN/C (complex nominals per clause) index recording larger values for the EFL group with small effects in results and conclusion sections. The English L1 group also produced more verb phrases (as indicated by VP/T or verb phrases per T-unit) in the abstracts, method, and conclusion sections with medium to large effects. The EFL group, on the other hand, produced more coordination structures, marked by CP/T (coordinate phrases per T-unit) and CP.C (coordinate phrases per clause) in the literature review and result sections.

Overall, these findings suggest that the English L1 students produced more subordination structures and lengthier sentences which are believed to mark higher-proficiency learners and the EFL group produced more coordination structures which are believed to mark lower-proficiency learners. For detailed discussions of these findings see 6.3. These findings have implications for EFL pedagogy that will be discussed in the final chapter.

B3. *To what extent do the ESL students who benefit studying in the UK academic setting perform better than their EFL peers who study English in a non-English-speaking context, and to what extent do the ESL students' performances approximate the English L1 group considering the effect of the shared academic setting (i.e., academic programmes, materials, syllabi, and immersion in an English-speaking academic context)? Do any such differences have implications for ESL academic immersion programmes?*

All things considered, the ESL group's performance in terms of producing the lexically and syntactically complex academic texts is generally very similar to the English L1 students, which means that they also outperformed the EFL group in terms of the values of various lexical and syntactic indices. There are some exceptions, though. For instance, in the literature review section, the EFL group outperformed the ESL group in most of the lexical diversity indices that showed between-group differences; the EFL group produced more verb and noun types in the method section as well as more lexically sophisticated words (as calculated via ls1) in method, results, and conclusion sections. Concerning the syntactic structures, the EFL group only produced more coordinated phrases (as measured via CP/C) than the ESL group, which as discussed earlier, points to the higher proficiency of the ESL group.

The differences between the English L1 and ESL groups are trivial (e.g., with small effect sizes) and not consistent across all rhetorical sections: the ESL group produced marginally longer sentences (marked with T-units), verb phrases and clauses only in the results section as well as more varied texts (calculated via mtld and msttr only) in the conclusion section. This comparable, and in many cases indistinguishable performance of the English L1 and ESL groups points to the possible effect of academic ESL immersion programmes and the role of shared materials, syllabi and academic contexts between these two groups in master's programmes in the UK. These findings are also consistent with the results of Bulté and Housen's (2014) study of intensive ESL academic writing programme and the ESL students' progress in syntactic complexity as evidenced by seven syntactic complexity measures as well as the subjective ratings of writing quality. The discussion of such implications will be presented at greater length in the final chapter. However, as will be mentioned in the next chapter, I did not have access to the proficiency levels of these students (e.g., a formal proficiency test) prior to the analysis of this corpus and therefore, I cannot make a definitive decision regarding the role of academic immersion programmes in the comparable statistics of the ESL and English L1 groups.

6.8.3. Answering Group C of Research Questions: Lexical and Syntactic Features of Postgraduate Academic Writing

This group of research questions with a rhetorical-based approach to academic writing deals with the prominent linguistic features (i.e., lexical and syntactic constructs) characterising each of the six sub-sections of master's dissertations (also called the sub-genres of the dissertation in this study). The sub-questions are formulated as:

C1. *What are the overall (dominant) lexical features of each of the six rhetorical sections of MA dissertations in terms of the lexical constructs of density, diversity and sophistication of the whole corpus? What are the top lexical predictors of each of the six rhetorical sections produced by all three groups combined?*

The rhetorical sections of method, result, and conclusion which are generally more reporting and descriptive appeared to be more lexically dense (particularly for the English L1 and ESL groups) than the rhetorical sections of introduction and literature review which are more explanatory and informational. All groups produced more lexically dense texts in literature review and result rhetorical sections. Regarding lexical diversity in the texts of the three groups, values of different lexical diversity measures are larger in introduction, conclusion and literature review sections. Finally, the rhetorical sections of introduction, method, and results are more lexically sophisticated than other sections.

The top overall lexical predictors of membership to all rhetorical sections are rttr, logttr and ls2 as indicated by graph 6.12. The first index, rttr or Root TTR, is the ratio of types to the square root of tokens, the second index is a logarithm-based measure of lexical diversity and the third one is a lexical sophistication index of the proportion of lexical types. This means that there were greater amounts of variance explained by these three indices which could better distinguish/predict texts belonging to any of the six rhetorical sections. Regarding the top lexical predictors of individual rhetorical sections, the rttr index can better classify abstracts, literature review, methods, and conclusion sections; the logttr index could better predict/classify introduction and result sections, and the ls2 measure is a good distinguisher of literature review, abstracts, and conclusion sections. These findings indicate that there are noticeable differences in various rhetorical sections regarding (the amounts/values of) lexical diversity and to a lesser extent lexical sophistication, especially in the literature review sections where more diverse and varied lexical words and types are used by all groups.

The lexical density of various rhetorical sections seems to be similar and the ld measure has the least predictive power for correctly assigning texts to their relevant rhetorical sections. This is in sharp contrast with the between-group differences and with the top predictors of group membership where the ld measure is found to distinguish between the three groups (and hence a good indicator of lexical complexity of groups of students with different English language backgrounds) as well as a strong predictor of group membership. These findings are also clear evidence to different effects of groups and rhetorical sections on the values of these lexical indices as well as the effectiveness of different lexical indices in explaining such effects and as indicators and predictors of proficiency and class membership.

C2. *What are the overall (dominant) syntactic features of each of the six rhetorical sections of MA dissertations in terms of the syntactic constructs of the length of production units, amount* *of subordination, amount of coordination, and degree of phrasal sophistication in the whole corpus? What are the top syntactic predictors of each of the six rhetorical sections produced by all three groups combined?*

The top three overall important syntactic structures across all rhetorical sections are CT/T (complex T-units per T-unit), DC/C (dependent clauses per clause), and CN/T (complex nominals per T-unit). This means that the two subordination indices, as well as the index marking phrasal sophistication, can better capture the variance among the six rhetorical sections and therefore, better predict and correctly classify any given text (in this study's corpus) into their relevant rhetorical sections.

With respect to individual sections, complex T-units are better predictors of literature review and result sections, dependent clauses are the noticeable predictors of the literature reviews and method sections, complex nominals (via both CN/T and CN/C combined) are shown to better predict method sections and literature reviews, and finally, coordinate phrases (via both CP/C and CP/T combined) could better classify results, literature reviews and abstracts. Relatively greater amounts of verb phrases were also produced by the three groups in the method and literature review sections.

The top syntactic predictors of rhetorical sections are quite similar to the top predictors of group membership. This finding is in obvious contrast with the relevant finding in the lexical dataset where the top lexical predictors of the group and rhetorical membership were dissimilar.

The three postgraduate groups have noticeably used greater amounts of all types of syntactic structures investigated in this study specifically in literature review and result sections; this suggests that these two rhetorical sections are characterised by various types of syntactic structures overall and that the three groups can better employ various syntactic structures in longer sections. On the other hand and across the three groups, the shorter texts such as abstracts and introduction rhetorical sections are mainly characterised by greater amounts of coordinate phrases compared to other indices (see graph 6.14).

6.8.4. Answering Group D of Research Questions. Statistical Modelling: Best-fitting and Most-accurate Models

The final category of research questions pertains to the regression and classification problems and predictive models. The linear mixed-effects models (regression problems) seek to predict the values of the lexical and syntactic indices given the classification labels of groups and rhetorical sections as the fixed effects. The classification problems, on the other hand, deal with predicting the correct classes/categories of these two variables given the values of the selected lexical and syntactic indices. For details of these statistics see sections 6.6 and 6.7. Accordingly, the following sub-questions are derived:

D1. *What are the effects of groups (English language background as English L1, EFL, and ESL) and rhetorical sections (the six sub-sections of MA dissertations), and their additive and interaction effects on the values of 22 lexical and 11 syntactic complexity indices? What are the best-fitting models which can explain the largest amounts of variations for these measures?*

Among the four models specified in 6.6, the model with the interaction effect of groups and rhetorical sections explains the greatest amounts of variation in the values of most lexical and syntactic complexity measures. This means that the value of a given index for any given group depends on the rhetorical function of the text (e.g., depends on the rhetorical section). The next-best model based on the model fit indices is the additive effects of groups and rhetorical sections (i.e., the arithmetic sum of the values of each of the fixed effects of groups and rhetorical sections), but the effect of groups on the value of an index does not depend on the effect of rhetorical section/type of text, in this case. Between the two remaining models which investigate the individual effects of either groups or rhetorical sections on the values of lexical and syntactic measures, the rhetorical-section-only model better captures the variations in both lexical and syntactic datasets, i.e., it explains most of the variability of that value around its mean.

D2. *How accurately can we classify the groups of students based on the values of 22 lexical and 11 syntactic indices obtained from the analysis of academic texts (all six rhetorical sections combined)? What are the specifications of the best predictive models of group membership?*

The results of the random forest predictive modelling for predicting accurate classifications of group membership based on the values of 22 lexical and 11 syntactic indices are presented in table 6.39 and discussed in 6.7.1. After the process of parameter tuning, the best-fitting lexical model obtained an accuracy of 54% with a CI of [40,67]%. This indicates that the random forest classifier in the caret package could obtain a maximum of 67% accuracy in predicting each text’s correct group on any unseen data (e.g., another sample with the same specifications). However, considering the similar performances of the ESL and English L1 groups as discussed in the analyses of variance in 6.3, I hypothesised that removing either of the ESL or English L1 groups from the model would result in a higher accuracy value and a lower classification error. This hypothesis was tested and the second model with the EFL and English L1 groups obtained 78% accuracy with a CI of [63,90]%. This finding suggests that the model is fundamentally a good model and the values of the 22 lexical indices contribute to a highly-accurate predictive model for correctly classifying each text to the relevant group of the student (English L1 vs. L2) who produced that text, i.e., these groups with different English language backgrounds produced overall different lexically dense, diverse, and sophisticated texts.

Likewise, the initial model based on the 11 syntactic indices obtained 51% accuracy with a CI of [38,64]% for correctly classifying texts in their relevant groups. Since the ESL and English L1 groups also performed very similar in terms of the values of different syntactic measures, a second model was specified by removing the ESL group; the accuracy of this new model jumped to 57% with a CI of [41,72]%. Once more, this significant gain in accuracy points to the fact that similar performances of the ESL and English L1 groups lowered the initial model’s accuracy and that the 11 syntactic indices investigated in this study are relatively good predictors of group membership overall. For the detailed discussion of these findings see section 6.7.1 and table 6.39.

D3. *How accurately can we classify each of the six rhetorical sections of MA dissertations in this study’s corpus based on the values of 22 lexical and 11 syntactic indices of the three groups of postgraduate students? What are the specifications of the best predictive models of membership to rhetorical sections?*

The random forest classifier in the caret package could predict the membership/classification of the six rhetorical sections based on the values of the 22 lexical complexity measures with an accuracy of 59% and a CI of [54,64]%. This shows that these lexical indices are good predictors for the classification of rhetorical sections in MA dissertations on any new sample with the same/similar specifications as the design of this study. As discussed in 6.3.1, the three groups produced similar amounts of certain lexical units in some rhetorical sections, which in turn resulted in somewhat lower classification accuracy by the classifier; this effect can be seen for instance in the literature review and method sections. Among the six rhetorical sections, the abstract section received a higher classification accuracy (93%) and the method section received the lowest (44%).

Notwithstanding a careful process of parameter tuning, the model which predicts the membership to rhetorical sections based on the values of 11 syntactic indices obtained 35% accuracy with a CI of [30,40]%. This finding suggests that the syntactic indices investigated in this study are reasonably good predictors for classifying the rhetorical sections, given the fact that this accuracy is well above the chance level. This model’s highest accuracy is recorded for the abstract section (56%) and the lowest for the conclusion section (21%). However, this finding also indicates that the three groups used similar amounts of syntactic structures of various types in different rhetorical sections (i.e., they maintained a similar overall style of writing in terms of syntactic structures throughout their dissertation). In other words, although these indices do not contribute to very high accuracy for the classification model of rhetorical sections, they are consistently produced throughout the dissertations with little variability across the three groups. It would be interesting to see if similar results are obtained from the investigation of different academic writing corpora, e.g., in different disciplines to examine the patterns of these syntactic measures in rhetorical sections of various disciplines, and/or to consider stylistic variations in this regard.

7 Concluding Remarks : Conclusions, Implications, and Suggestions for Future Research

7.1. Overview and a Brief Summary of the Research

This chapter begins with a summary of this study, and proceeds with a review of the main findings and a detailed explanation of the conclusions and implications of the findings including research and methodological implications. In the next sections, I will discuss the limitations and delimitations of this project and how future researchers could address these limitations along with the specific recommendations of this study to direct informed research projects.

This research is an interdisciplinary study that adopted the principles of corpus linguistics and the methods of statistical modelling to analyse the rhetorical sections (i.e., the six sub-sections of MA dissertations) written by postgraduate students with different English language backgrounds. This study had a four-fold purpose corresponding to the four categories of research questions discussed in 6.3.1.

First, through a quantitative measure-testing process, the efficacy of each of the 22 lexical and 11 syntactic complexity measures in capturing differences of L1 vs. L2 academic texts was examined, the relationship between and among these measures and their relevant constructs was investigated, the structure of the categories of the constructs and measures was tested against the current classifications in the literature and further explored, and finally, both overall and specific indicators and predictors of linguistic proficiency were determined to assist future studies with the measure-selection process and for a more expansive picture of the efficacy of the selected measures as indicators and predictors of lexical and syntactic proficiency.

The second aim was to compare the lexical and syntactic performances of the three postgraduate groups with different English language backgrounds of EFL, ESL, and English L1 in each of the six rhetorical sections and to find the overall and specific linguistic complexity differences.

The third objective was to find out the prominent linguistic features (e.g., lexical and syntactic constructs) of each of the six rhetorical sections conventional of a dissertation based on consistent patterns that had emerged in this analysis of MA dissertations of various sub-disciplines of applied linguistics.

The final aim was to investigate the usefulness of three statistical modelling methods (structural equation models, linear mixed-effects regression-based models and machine learning random forest classification models) to examine the structure of the data and to build predictive models of lexical and syntactic complexity in postgraduate academic writing. The mixed-effects models examined the effects of one text-extrinsic factor (i.e., groups of students with different English language backgrounds and academic contexts) and one text-intrinsic factor (i.e., the rhetorical sections of dissertations with various communicative purposes) on the values of lexical and syntactic complexity indices and the amount of variation that each type of model can explain for each index. The random forest models, on the other hand, examined how well we can classify the texts into their relevant groups and rhetorical sections given the values of the lexical and syntactic measures and what the top lexical and syntactic predictors of linguistic proficiency are.

7.2. Conclusions and Implications of this Study

As explained in the previous section, the multifaceted nature of this research called for various types of tests, each gauging this study's corpus from a different angle. This diversity leads to a number of conclusions and implications that will be discussed in the following sections. In sections 7.2.1 to 7.2.4, I first reiterate the main findings and conclusions and then discuss the implications in light of previous studies and recommendations of other researchers in the field.

7.2.1. Implications for Corpus-based Research on Rhetorical Sections and Linguistic Features (Indicators and Predictors) of Academic Writing Proficiency

Despite the host of genre analysis studies which focus on the genre moves and other rhetorical functions/expectations of various types of texts, insufficient and sparse attempts have been made to identify predominant lexical and syntactic characteristics and structures of the main rhetorical sections of various types of academic writing, especially theses and dissertations for specific disciplines (e.g., Hinkel, 2003; Hyland & Tse, 2005; Jalali & Ghayoomi, 2010; Thompson, 2002). This is particularly important for English L2 postgraduates who, not only have to submit a substantial piece of academic writing in the form of a thesis/dissertation, but may be required to produce high-quality journal articles and other types of academic texts to sustain their academic success. Large-scale corpus-based studies of the identification of predominant lexical and syntactic features and structures of rhetorical sections produced by English L1 postgraduates and professional academics can offer an insight into the required proficiency features and characteristics of different rhetorical sections for English L2 students. They may also serve as effective (teaching) guide for EAP practitioners and materials developers in EFL academic contexts. Identification of the predominant lexico-grammatical features of various written registers (e.g., academic writing) has already been conducted by Biber and Gray (2013) and their other similar works, but they did not drill down into the rhetorical sections to specify the linguistic features characteristic of each sub-genre.

This study's findings point to some prominent lexical and syntactic structures and constructs that characterise the rhetorical sections of a dissertation or thesis that could be addressed in thesis writing modules, especially in EFL academic settings. Along this line, Hinkel (2003) believes that the pedagogical implications of such findings are to "bring learners' attention to issues of divergent L2 registers and genres, focus on syntactic and lexical manifestations of various registers and genres in text, and emphasize the importance of appropriate grammar and lexical range in written academic text" (p. 281). Applying the results of corpus research to L2 pedagogy is stressed by Yoon and Hirvela (2004) as well. They emphasise the role of "corpus pedagogy" and meaningful input as well as genre-based corpus analysis in L2 and EAP writing instruction to familiarise the students with genre-specific expectations of linguistic patterns for achieving "high levels of proficiency as L2 writers" (p.259). In recent years, the necessity of conducting rhetorical and context-based investigations of linguistic complexity measures is felt more than before because of this realisation than certain lexical and syntactic features/structures surface on certain types of texts (e.g., based on the rhetorical functions and genre expectations) than others. Lu (2017) for instance, calls for the investigation of syntactic complexity measures based on the genre, context, and the proficiency level of L2 writers. The idea behind such analysis, as Flowerdew (2017) emphasises, is that "a given lexical and grammatical item must be related to its particular rhetorical purpose" and that "these purposes vary according to register" (p. 92) which should be the focus of writing for specific purposes and EAP writing instructors. Along the same line and commenting on lexical diversity, McCarthy and Jarvis (2010, p. 382) emphasise that "different rhetorical purposes and strategies may necessitate that different parts of a text have different diversity levels". In this study, for instance, by investigating a corpus of dissertations based on various rhetorical sections, I demonstrated throughout chapter 6, and specifically in the answers to the research questions in group C, how certain lexical and syntactic constructs surface more in some of the rhetorical sections with different length, rhetorical expectations, and written by students with different English language backgrounds. The results show, for instance, that the introduction sections are more lexically diverse and sophisticated and that literature reviews feature more subordination (especially the use of dependent clauses) and phrasal complexity (especially the use of complex nominals). The pedagogical implication of this is ‘not to’ teach linguistic features (e.g., lexical and syntactic constructs) in isolation, but ‘to’ put them into perspective, e.g., based on the disciplinary and genre expectations. Apart from the implications for ESL and EAP programmes, these results also have implications for writing research, specifically genre-specific and discipline-specific research on academic writing.

Discipline-specific academic writing research and instruction is the focus of another line of research on EAP that gained momentum with the works of Ken Hyland, and John Flowerdew among others. Hyland and Tse (2007), Nation (2013), Durrant (2014), and Coxhead (2018) for instance, demonstrated that vocabulary use is different across disciplines, e.g., pointing to the use of technical vocabulary in Medicine and Botany, specialised adjectives in Philosophy, specialised nouns in engineering, and specialised verbs in Science. This, in turn, results in different lexical and syntactic patterns in the texts of linguistics-related disciplines (like this study) compared to other disciplines, that can be best investigated via large-scale corpus-based discipline-specific studies and implemented via explicit instruction, especially for English L2 postgraduate students. One of the few systematic analyses of disciplinary variation of linguistic features is the work of Green (2019) which examines a large set of lexical sophistication, and syntactic sophistication and complexity indices across eight disciplines belonging to humanities, social sciences, and hard sciences.

Even though a general-purpose list of academic vocabulary benefits the students in various disciplines and proficiency levels, discipline-specific vocabulary and the specialised language of a discipline, as Coxhead (2018) and Woodward-Kron (2008) state, contribute to disciplinary knowledge for members of the same discourse community or as Bloch (2008) calls, a shift towards the ‘local knowledge’ of language. An example of unawareness of such discipline-specific vocabulary can be found in the linguistic excerpt from the EFL text that was presented earlier in chapter six:

“When EFL learners say ‘tree’ instead of ‘three’, they should not expect the native listener to get what they have wanted to produce at the first step because the addressee does not live on their mind.”

In this sample sentence, the use of more discipline-specific vocabulary and phrases such as ‘mispronunciation’, ‘treating the two words as homophones’, ‘phonetic difficulties’, ‘difficulties in pronouncing digraphs’, ‘voiced [th] vs. unvoiced [t] sounds’, and ‘misinterpretation by the English L1 listener’ could have enhanced the quality of the writing and contributed to a more lexically-sophisticated text overall.

As reflected in the discussions in 6.3.1 and 6.3.2, most discipline-specific vocabularies are also considered as sophisticated words because of their absence in the frequently-used word lists. Upon manual inspection of the data, I found that the use of discipline-specific vocabulary in the EFL texts are more pronounced in the method, result, and conclusion rhetorical sections mainly where the students paraphrase experts’ opinions using the experts’ choice of words. I noticed underuse of such discipline-specific words in the introduction sections of EFL texts mainly where the students tend to explain their own research in their own words and/or justify the significance of their own study, etc. This, as elaborated in chapter six, could be one main reason for the disparities between the values of ls1 and ls2 sophistication indices as well.

In recent years, other scholars have also pushed this agenda in what Friginal (2013) states as the necessity of discipline-specific and genre-specific corpus-based research for identifying linguistic features (especially vocabulary and grammatical features) and characteristics of academic writing produced by students with different proficiency levels to aid teaching of writing for specific purposes across various disciplines. In the present study, I collected MA dissertations from various sub-disciplines of applied linguistics and therefore, the findings have more relevance and applicability for researchers and EAP practitioners in these fields.

7.2.2. Implications for Measure Selection and Evaluation and an Application for L2/EAP Writing Assessment

The multilayered task of evaluating 22 lexical and 11 syntactic complexity measures in terms of their performance, the overall and specific structures, efficacy in capturing the proficiency differences, and usefulness in predicting group and rhetorical-section classification culminated in multiple implications for measure selection, testing and evaluation particularly for L2 and EAP writing assessment. Since the detailed discussion of the findings of these processes have been presented throughout chapter six, here I will provide brief summaries of conclusions that were drawn from each analysis, and focus on the implications on the evaluation processes, construct-distinctiveness, the efficacy of measures in each pair/group of similarly-calculated measures, and their effectiveness in predicting linguistic complexity (differences) and in classification problems.

The evaluation of lexical and syntactic complexity indices for selecting relevant and effective measures for assessing academic writing is an arduous and unforgiving task due to the sheer number of lexical and syntactic indices that are proposed and used in the literature. The TAASSC programme (Kyle & Crossley, 2017) for computing syntactic complexity and sophistication, for instance, offers 372 indices that are reported in the literature as useful measures for various English linguistics studies. Other programmes compute a similarly-large number of indices: TAALES (Kyle & Crossley, 2016) for analysing lexical sophistication with 484 measures, CRAT (Kyle & Crossley, 2016) for lexical sophistication and cohesion with 700 indices, and the lighter programmes like Coh-Metrix (Graesser et al., 2004) for assessing writing cohesion, readability which includes syntactic and lexical indices with over 200 measures, CTAP (Chen & Meurers, 2016) with 180 lexical and syntactic complexity indices, and the L2SCA analyser (Lu, 2010) for analysing syntactic complexity with 14 indices, etc. Screening the findings of studies in the area of linguistic complexity is not without its complications either. The multiplicity of studies with different research designs**,** objectives, corpora, variables, and groups of learners makes the selection of a set of relevant and effective indices for the specific objectives of a study a formidable task. In the case of the present study, finding sets of lexical and syntactic complexity measures which have shown to be reliable in capturing the proficiency differences of postgraduate writers was challenging because of the scarcity of previous related works on postgraduate academic writing, especially thesis/dissertation writing.

For these reasons, I started with a large list of measures and narrowed them down in several stages to obtain a short list of indices whose performance were shown to fit this study's objectives as closely as possible. These two sets of indices were then subject to a pilot study to examine their effectiveness further with a subset of texts collected up to that point. Some measures were dropped at this stage as well to arrive at the final sets of measures that showed between-group differences for at least one pair of comparison. Having concluded this study, I presented the findings of various statistical tests to recommend a group of the most effective measures for each type of analysis, e.g., the indices that effectively captured between-group differences, lexical and syntactic indicators and predictors of proficiency, and the top predictors of rhetorical sections and groups classification with implications for measure selection of future related studies.

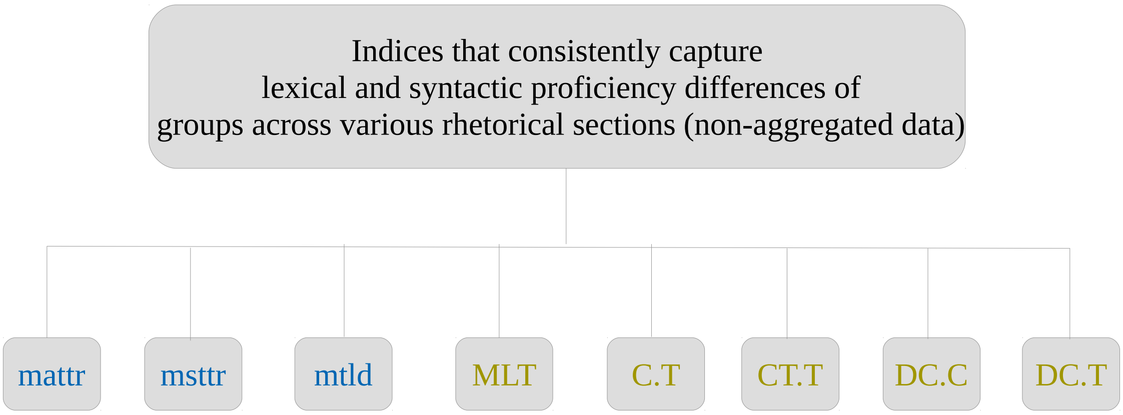
The next point in the measure selection process is the construct-distinctiveness that was examined via the two statistical methods of correlation analysis and confirmatory and exploratory factor analyses.

The correlation tests (section 6.4) in both lexical and syntactic datasets revealed that measures that are commonly reported/assigned to the overall lexical and syntactic categories/constructs have stronger correlations with each other than with the indices in other constructs. Clear boundaries (e.g., the mean values of indices) between the main lexical and syntactic constructs corroborate these findings further. This is the first time that the extended sets of 22 lexical and 11 syntactic complexity measures have been tested for construct-distinctiveness in a postgraduate academic writing corpus of MA dissertations. Lu (2012) and Šišková (2012) have previously confirmed the construct-distinctiveness of some of these lexical indices in a corpus of oral and written narratives respectively. Even though some of the measures used in this study were not investigated in these two studies, the overall distinct nature of the constructs of lexical density, diversity, and sophistication in all studies point to the fact that this construct-distinctiveness is indeed independent of the mode of language. Interestingly, this pattern is also observed in the sub-constructs of lexical diversity of logarithm-based indices, word-string-based measures, and indices based on the type-token ratio of word classes where measures assigned to each sub-construct showed stronger correlations with each other than with the indices in other sub-constructs (tables 6.21 and 6.22). Similar results were obtained in regarding the overall constructs of syntactic coordination, subordination, and phrasal complexity.

Similarly, the results of confirmatory and exploratory factor analyses (CFA and EFA) confirmed the overall lexical and syntactic constructs/categories reported in the literature, except for some misplaced measures. As explained in the discussion of findings, CFA is sensitive to outliers (e.g., the assumption of multivariate normality), multicollinearity across measures of different constructs, large numbers of variables, sample size, and even the presence of a few nonsymmetrically-distributed residuals of covariances (see the discussions in Tabachnik & Fidell, 2013, pages 730-739 the discussion of Ullman). These are some reasons that the models of constructs and their representative measures suggested in the literature did not produce acceptable fit indices with this study's corpus of MA dissertations. As a result, follow-up EFA tests were carried out to find out the actual structure of the constructs with their assigned measures based on the corpus of this study and to locate the misplaced indices that caused the CFA tests to produce unacceptable fit indices. In EFA tests new sub-constructs were detected in the lexical dataset, and a few misplaced measures were observed in both lexical (e.g., ndwerz, cvv1, rttr, vv1) and syntactic (e.g., MLT, VP.T) datasets; however, both datasets revealed structures that correspond to the overall categories/constructs present in the literature.

A summary of the more effective indices based on the results of various statistical tests in chapter six is presented in diagrams 7.1 to 7.3. Diagram 7.1 shows lexical and syntactic complexity measures that could consistently capture between-group differences across all six rhetorical sections (results derived from tables 6.5 to 6.11). As demonstrated in diagram 7.1, lexical indices as indicators of lexical proficiency differences are all from the category of lexical diversity based on word strings/segments and syntactic indices as indicators of syntactic proficiency are mainly subordination measures. More discussions on this have been provided in chapter six, section 6.3, as well as in the answer to the research question A3.

Diagram 7.1. Effective measures for capturing lexical and syntactic complexity differences


Indices that consistently capture
lexical and syntactic proficiency differences of
groups across various rhetorical sections (non-aggregated data)
mattr
msttr
mtld
MLT
C.T
CT.T
DC.C
DC.T

mattr: moving-average TTR
msttr: mean segmental TTR

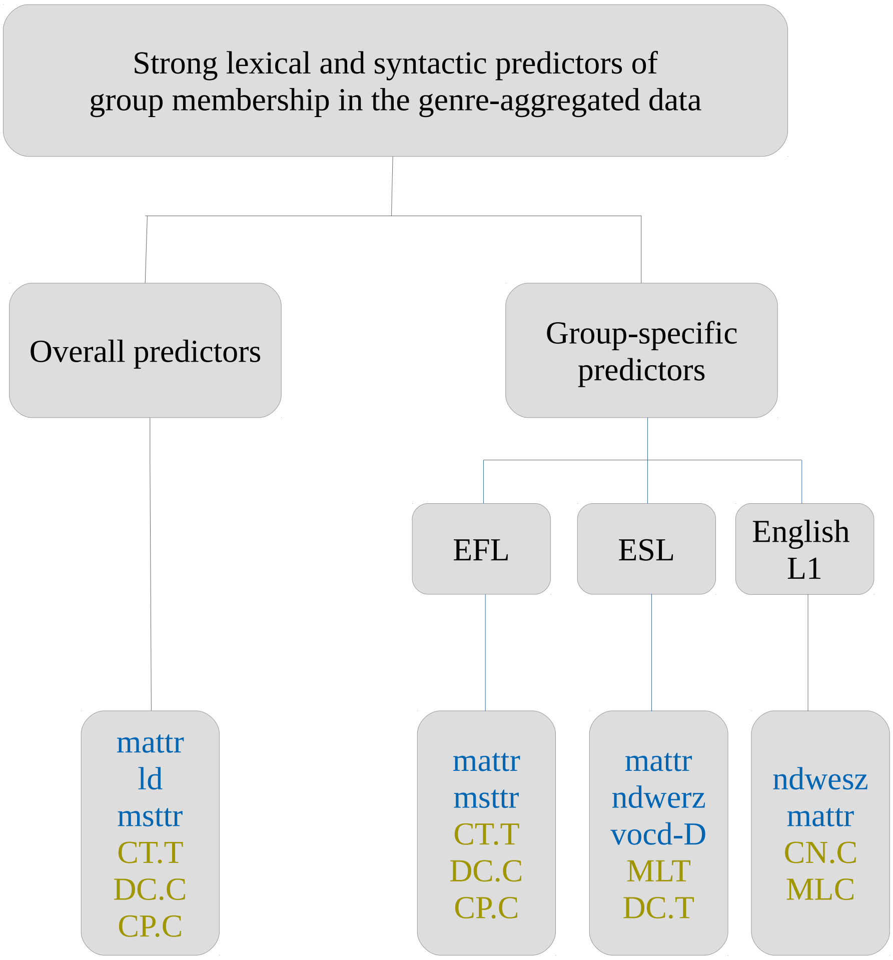
mtld: measure of textual lexical diversity

MLT: mean length of T-unit

C.T: clauses per T-unit
CT.C: complex T-units per T-unit
DC.C: dependent clauses per clause
DC.T: dependent clauses per T-unit

Diagram 7.2 shows the overall and group-specific indices as predictors of group membership in the genre-aggregated data (the entire dissertations: all six rhetorical sections collapsed as a weighted mean). This is to predict which lexical and syntactic complexity measures could more accurately classify the groups of students (with different English language backgrounds) based on the values of these indices as obtained from the analyses of the entire dissertations. As is demonstrated in this diagram and discussed in more detail in section 6.7.1, the lexical predictors are mainly lexical diversity of word strings/segments and the original D measure as calculated via the vocd-D index. Lexical density is also shown as a strong overall predictor of group membership. Two syntactic subordination indices as well the coordination index of CP.C are also shown as strong predictors. Each of the measures in each category are suggested as strong predictors of group membership across postgraduate groups in postgraduate academic writing, especially theses and dissertations and by extension in research articles or research reports. The reliability of these indices as predictors of group membership in research articles by expert writers could further be investigated in future studies.

Diagram 7.2. Strong lexical and syntactic predictors of group membership in postgraduate academic writing



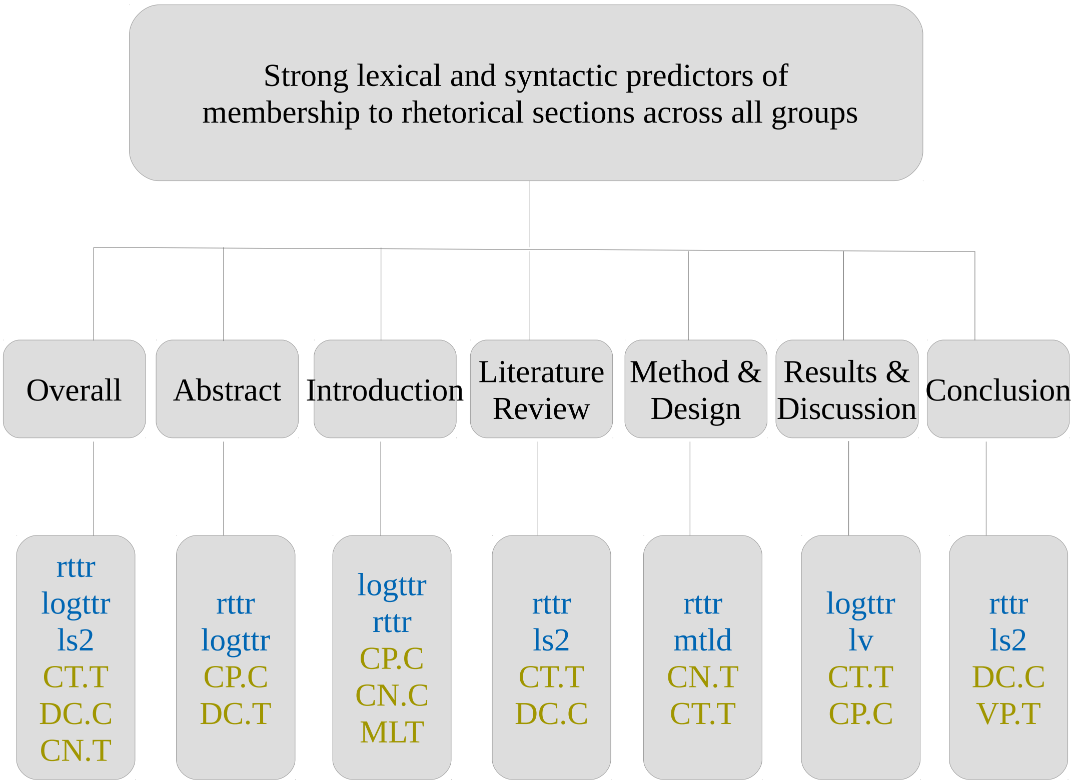
mattr: moving-average TTR
ld: lexical density
msttr: mean segmental TTR
ndwerz & ndwesz: variations of number of different words
Vocd-D: original D measure

CT.T: complex T-units/T-unit
DC.T: dependent clause /T-unit
DC.C: dependent clauses/clause
CP.C: coordinate phrases/clause
CN.C: complex nominals/clause
MLC: mean length of clause

Diagram 7.3 demonstrates the overall and rhetorical-section-specific lexical and syntactic measures as strong predictors of membership to each rhetorical section for all groups. This is to predict which of the lexical and syntactic complexity measures can best classify the six rhetorical sections (labelled as ‘genre’ in the tables and graphs in this thesis) given the values of these indices as obtained from the analyses of dissertation sections produced by all three groups of students. As is illustrated in this diagram and discussed in detail in section 6.7.2, and unlike the measures in the previous two diagrams, the lexical diversity measures of rttr and logttr as well as the lexical sophistication of ls2 are shown to be strong predictors of the classes of rhetorical sections of master’s dissertations as distinct sub-genres. This is an evidence of the distinct nature of text classification based on text-extrinsic factors (e.g., the groups of students) vs. text-intrinsic factors (e.g., the features of texts in terms of rhetorical sections and coomunicative goals) regarding the use of lexical indices. However, the two subordination indices in this diagram were also shown as overall strong predictors of group membership in diagram 7.2 as well. These results are similar to the results of mixed-effect models in which rttr as well as CT/T and DC/T indices showed largest variations in most models that gauged the interaction effects of groups and rhetorical sections as discussed in detail in sections 6.6.2 and 6.7.2.

An implication of these findings is that future researchers who need to restrict the number of indices in their assessment of writing, especially in assessing writing proficiency of postgraduates, could use the most effective measure in each construct and sub-construct based on the results of this study and based on the objectives of their project (e.g., finding proficiency differences or predicting the quality of texts and/or classification of variables). Another implication of these findings and a take-away message is that, even though some studies (including this study) found that the categories of the lexical and syntactic constructs and sub-constructs correspond, specific differences could be spotted based on the type of the data that is investigated, e.g., the genre of writing, the presence of specific rhetorical sections, writer differences in terms of the English language background, and the sample size. Therefore, a methodologically more appropriate approach is to choose the measures that can consistently capture texts’ complexity differences and the measures that tend to line up on the same factor/construct in different studies (e.g., studies with written/oral data, varying sample sizes, learner language backgrounds, different genres, general English vs. academic writing, etc).

Diagram 7.3. Strong predictors of genre/rhetorical-section membership in postgraduate academic writing



rttr: root TTR
logttr: logarithmic TTR
ls2: lexical sohphistication type II
lv: lexical variation
mtld: measure of textual lexical diversity
MLT: mean length of T-unit

CT.T: complex T-units/T-unit
DC.C: dependent clauses/clause
DC.T: dependent clauses/T-unit
CN.T: complex nominals/T-unit
CP.C: coordinate phrases/clause
VP.T: verb phrases/T-unit

A further contribution of this study is the additional development of the LCA analyser in the form of LCA-AW programme (version 2.1; Nasseri & Lu, 2019) for investigating the lexical sophistication indices in academic writing texts in applied linguistics and language studies. LCA-AW is a modified version of the LCA programme for analysing lexical complexity indices that was built by Xiaofei Lu (2012) and includes 25 measures belonging to the constructs of lexical density, diversity/variation, and sophistication as explained in chapter five. For calculating lexical sophistication indices, LCA uses the BNC general English word list as the reference point to filter the words that do not appear in the top 2000 frequently-used words in this word list. This programme was designed based on the advanced texts in the first and second language acquisition and development literature and not academic texts specifically. Academic texts often include frequently-used specialised words or terminology that are specific to any discipline. Therefore, I included the BAWE corpus word list for linguistics and language studies as another reference point besides the BNC word list. This means that in LCA-AW, the lexical sophistication indices outlined in chapter five will be computed as those that do not appear in the first 2000 frequently-used words in the BNC corpus (with an alternative of using ANC or American National Corpus) as well as those that do not appear in this BAWE word list. Future researchers who wish to investigate these lexical sophistication indices in various academic writing genres, rhetorical sections and across various sub-disciplines of linguistics could use this open-source and free programme written in Python; further details of this programme and the ways to download and use it are presented in Appendix D.

7.2.3. Methodological Implications for Building Statistical Models of Linguistic Features and Indices

A contribution of this study is the proposal of the use of structural factor analysis and the statistical modelling methods of linear mixed-effects modelling and predictive modelling based on a machine learning approach e.g., random forest or gradient boosting for a more expansive investigation of linguistic indices and to examine the effectiveness of and the relationships between the indices from different perspectives.

Structural factor analysis recommended by McArdle (2011), is a powerful theory/hypothesis testing and/or validation technique that is more prevalent in social sciences studies that use applied multivariate statistical analysis with multiple variables. This structural method includes both confirmatory and exploratory analyses: the former method tests the accuracy of previously-proposed models (usually theoretically-driven models) and the latter method allows for theory development and the detection of relationships between variables in a particular dataset (for detailed discussions on SEM methods see Byrne, 2006; Hershberger, Marcoulides, & Parramore, 2003; Kline, 2005; McArdle, 2011). The popularity of this method in behavioural and social sciences is mainly due to the importance of latent variables (e.g., constructs) that cannot be explicitly or precisely measured. They are usually quantified using the features or indices that operationally define them. Despite its strengths and potentials in examining the plausibility of hypotheses, confirming/rejecting proposed models in the literature, understanding the relationship between constructs and their observed indicators, and detecting new dimensions/constructs and structures, this method is only infrequently used in corpus-based research on linguistic features (see for instance the discussions in In’nami & Koizumi, 2011 on its application in language testing and learning research). Relevant studies have used this method mainly to find the effects of multiple independent variables on language development and learners’ test scores (see for instance the review of literature in In’nami & Koizumi, 2011).

In the present study, factor analyses were used to verify and examine the lexical and syntactic classifications proposed in the literature (e.g., the structure of lexical complexity constructs and indices in Lu, 2012 and the structure of syntactic complexity constructs and indices in Lu & Ai, 2015) to test whether these classifications are supported with this study’s corpus of MA dissertations, to find the relationship between various lexical and syntactic constructs and their representative measures (e.g., the results of factor correlations and factor loadings), to detect new structures/dimensions in the lexical and syntactic datasets based on an specialised academic writing corpus, and to find the indices that best represent each construct/factor and the indices that are misplaced compared to the models proposed in the literature.

The second facet in the triad of statistical modelling in this study is linear mixed-effects models which elucidate the effects of one main text-extrinsic variable (i.e., groups of students with different English language backgrounds and academic contexts) and one main text-intrinsic variable (i.e., rhetorical sections of dissertations with various commnicative purposes) on the values of dependent variables (e.g., the 22 lexical and 11 syntactic measures in this study). These models also detect the models that explain the largest amounts of variation for each measure (e.g., the most important indicators of proficiency and/or the most effective indices representing a construct and/or the best-performing indices in capturing variation in the data). A strength of this type of modelling is that multiple models can be defined in parallel in the model specification stage to investigate the effects of each independent variable as well as the additive and interaction effects of such variables on explaining the dependent variable’s values. Care needs to be taken to compensate for the possible influence of spurious positives due to multiple significant tests, for example by adjusting the alpha level before interpreting the results. Another strength of this type of modelling compared to general linear models, is the versatility of considering the role of random effects on the values of dependent variables, for instance, the role of by-subject variability and multiple responses per student for multiple rhetorical sections, in this study’s case. Despite these advantages of mixed-effects models in multiple model specification and in finding the specific effects of each variable or a combination of variables on the values of linguistic indices, these tests are also infrequently used in the corpus-based research (compared to psycholinguistics) on investigating the patterns of various linguistic features. Gries (2015) specifies this type of modelling as the most under-used statistical method in corpus linguistics. For the usefulness and applicability of these methods and the ways to incorporate them in linguistics research design see discussions in Barth and Kapatsinski (2018), Cunnings and Finlayson (2015), Gries (2015), and Winter (2019).

The third statistical modelling method adopted in this study is the supervised machine learning method of random forest for predictive classification modelling. During the past decade machine learning methods, especially the tree-based method of random forest gained more popularity and attention in hard and soft sciences due to its flexibility in handling non-normally distributed residuals (e.g., by using bootstrapping methods) and its robustness to outliers, its needlessness of data standardisation (e.g., because of varying data scales across variables), its robustness to multicollinearity (e.g., correlated predictors) by using the feature bagging, its usefulness in solving both regression and classification problems (i.e., the CART models), and its capability in handling non-linear relationships (see for instance the discussions in Boulesteix et al., 2012; Probst et al., 2019; Shalev-Shwartz & Ben-David, 2014; Strobl, Malley, & Tutz, 2009; Ziegler & König, 2014 and the relevant discussions in various linguistics research in Baumann and Winter, 2018; Brown et al., 2014; Gries, 2019; and Thompson, Hunston, Murakami, and Vajn, 2017 among others). The two main functions of random forests are their effectiveness in predictive classification modelling (eg., the use of robust methods in classifying the quantitative values to their appropriate respective classes/categories) and the specification of the most effective/strong predictors of accurate classification among a relatively large set of variables.

In this study, this method was used to determine how accurately we can predict/classify the three groups of postgraduate writers (EFL, ESL, and English L1) and the rhetorical sections of dissertations based on the values of the 22 lexical and 11 syntactic indices and to determine the top predictors of lexical and syntactic proficiency (e.g., by way of variable importance features in random forest algorithms) on any unseen data. Apart from classification, this line of research has implications for automatic writer identification (e.g., using NLP tools) based on their English language background; similar research like Ai and Lu (2015) as well as Jarvis and Crossley (2012) have already incorporated the students' L1s into such research and emphasised the implications of these findings for automatic native language identification using NLP tools. These two lines of research, namely automatic identification and classification of writers and texts based on their L1s and English language backgrounds are significantly facilitated by using prominent linguistic features that can descriminate texts written by the afore-mentioned writers as well as, for instance, improving the performance of such systems using top lexical and syntactic predictors of group membership. The discussion of findings in section 6.7 indicates the degree of usefulness of including the groups of students as well as the rhetorical sections of dissertations as independent variables in future corpus-based research on linguistic features and/or indices as well as some guides on the parameter tuning methods for random forest algorithms for future studies with similar research design as this study.

The combination of these three strong statistical modelling methods renders a more expansive investigation of linguistic features and/or indices (e.g., lexical and syntactic measures) and a better picture of their effectiveness as indicators and predictors of (lexical and syntactic) proficiency in corpus-based research on writing (especially academic writing) that incorporates multiple linguistic indices or features as well as multiple factors/independent variables. Furthermore, this study benefited from such a unique perspective for investigating the role of groups and rhetorical sections in predicting the values of an extended set of lexical and syntactic complexity measures in a corpus of MA dissertations written by English L1 and L2 students in two different educational contexts.

Apart from the methodological implications of using these three distinct statistical modelling methods for investigating linguistic indices, this study offers a detailed guide on corpus construction, text preparation and cleaning processes and the ways to tackle some unwanted textual and non-textual elements (e.g., quoted texts, numerical values, notations, certain punctuations, typographical mistakes, symbols, URLs, contracted forms, and so forth as discussed in detail in 5.2.4), as well as text pre-processing and the ways to make the tokenisation, POS tagging, and lemmatisation processes consistent across several platforms/programmes.

7.3. Merging the Views of Linguistic Complexity and Some Suggestions for Academic Writing Instruction

Notwithstanding the detailed discussions on two main views of linguistic complexity and their relationships with linguistic proficiency in chapters one and three, it seems necessary at this stage that I bring all these points together for a more expansive outlook. The first point that is a central theme is how the two views of linguistic complexity in general, and lexical and syntactic complexity specifically, help to interpret the findings of this study and to show how the findings are related to linguistic proficiency.

By adopting the system view I have demonstrated how various lexical and syntactic complexity constructs and measures are inter-related and structured as a complex system where generally larger values across these measures lead to more complex academic texts, thus a higher overall linguistic proficiency. This point is reflected in both quantitative and qualitative analyses. The results of analyses of variance and between-group differences in the six rhetorical sections demonstrated that overall the English L1 group (followed closely by the ESL group) produced more lexically and syntactically complex texts and that the EFL group produced the least linguistically-complex texts regarding the values of most of the lexical and syntactic indices in most rhetorical sections. Since most of these indices have been previously confirmed by various researchers as indicators and predictors of proficiency (e.g., as gauged by holistic ratings, programme-based proficiency levels, and global-level proficiency levels such as CEFR), consistent and significant differences in the values of the reported complexity measures in this study can be attributed to proficiency differences.

Qualitative analyses of the excerpts from dissertations in 6.3.2 and 6.3.4 also reflect these differences mainly between the English L1 (followed by the ESL) and EFL texts regarding the constructs of lexical density, diversity, and sophistication as well as syntactic subordination and phrasal complexity. One English L1 excerpt, for instance, showed a compressed style of writing (i.e., lexical density) and a larger number of general and discipline-specific sophisticated items that are dispersed more proportionately throughout the text compared to the EFL excerpt that made use of a far smaller number of sophisticated words which are sparse. The EFL excerpt also showed a lower number of unique types and a writing style that did not include efficient use of discipline-specific vocabulary to convey the message. The ESL text exhibited even more number of varied lexical items, a more condensed style, but fewer sophisticated items compared to the English L1 text. I have also explained the quantitative results of lexical sophistication measures, in particular the findings of ls1 and ls2 for group differences in 6.3.1, and demonstrated some sample sentences from the EFL group from the relevant rhetorical sections. I have also explained in detail in the literature review, section 2.2.5, how higher lexical diversity can be achieved without the use of sophisticated lexical items, and vice versa.

Numerous researchers (e.g., Crossley & McNamara, 2012; Hinkel, 2003; Laufer & Nation, 1995; Lu, 2012) emphasised that the use of low-frequency and rare/unique/sophisticated words is considered to be a marker of a broad range of vocabulary and a reliable predictor of writing proficiency and the overall quality of academic writing. Hinkel (2003), for instance, observed that English L2 students with “a relatively high academic standing" "employ excessively simple syntactic and lexical constructions" and simpler linguistic structures than their English L1 peers. He also observed that the academic production of these advanced-level English L2 students predominantly consists of a small range of grammar as well as high-frequency and everyday vocabulary items. He maintains that in large-scale academic texts produced by English L2 students, "syntactic and lexical simplicity is often considered to be a severe handicap" (p. 275-276). Other researchers (e.g., Pica, 1985; Pienemann, 1985; Shahriari et al., 2017; Silva, 1993) also share these concerns and stress the necessity of explicit instruction for expanding English L2 students' lexical and syntactic repertoire, and as Hinkel (2003) phrase it, "to yield more sophisticated syntactic constructions and lexis so that the students are at a smaller disadvantage when they leave the ESL classroom" (p. 299).

Using varied and non-repetitive words as measured by lexical diversity indices is regarded by many scholars as an indication of proficient writers with wide active vocabulary knowledge (e.g., Housen et al., 2008; Kim, 2014; Lu, 2012; Read, 2000). The results of this study's between-group differences in most of the lexical diversity indices as demonstrated in tables 6.7 to 6.13, substantiates that the EFL group is not as competent as the other two groups regarding the diversification of lexis. This is also reflected in the analysis of sample texts from the dissertations. The EFL excerpt in 6.3.2, for instance, exhibits far smaller values of both lexical variation (lv) and verb variation (cvv1) than ESL and English L1 texts. More frequent use of function words and general-purpose words such as 'seems' and 'like' in the EFL excerpt has also been discussed. Since lexical diversity indices in this study capture this variation from different angles/with different formulas, the gap between the EFL and their ESL and English L1 peers calls for the EAP practitioners in Iran and other EFL contexts to include the concept of lexical diversity in the materials design process and to raise awareness about its effects on lexical proficiency.

Similarly, a lexically dense text, especially an abstract includes more informative words (i.e., the content words) which is crucial given the fact that almost all academic abstracts have a word limit and therefore, it is necessary for EAP practitioners in Iran and perhaps, in other EFL academic settings as well, to teach strategies to use this limited space to convey all relevant information. This could be achieved, for example, by using nominalisations, longer/elaborated noun phrases, and appositive noun phrases (see e.g., Biber, 2006; Biber & Gray, 2010, 2016). This overall 'compressed' characteristic of academic writing, as Biber and Gray (2010) emphasise, is crucial for expert readers to "quickly extract large amounts of information from relatively short, condensed texts" (p. 2). The sample texts from three groups' dissertations in 6.3.2 vividly show these differences in lexical density between the EFL and the other two groups, both as the ld values suggest and in the more frequent use of function words by the EFL student.

The ESL followed by the English L1 excerpts were also syntactically more complex than the EFL text regarding the use of subordination and phrasal complexity constructs, which, as discussed, are associated with higher syntactic proficiency and maturity. The EFL excerpt, on the other hand, showed more coordination (as coordinated phrases) that is believed to be a feature of lower levels of English proficiency compared to expert writers. The subordination differences are more noticeable with finite dependent clauses (labelled as 'DC' in this thesis). Even though this project did not quantitatively examine non-finite dependent clauses, the analysis of the sample texts from the dissertations of the three groups showed far fewer instances of non-finite subordinate clauses than finite ones. This is in contrast to the findings of some previous studies that show non-finite dependent clauses as a characteristic of (advanced) academic writing (e.g., in Biber and Gray, 2010; Biber et al., 2011; Staples et al., 2016). Staples et al. (2016), for example, provide evidence that the developmental sequence moves from finite dependent clauses (also as a characteristic of spoken discourse) to non-finite ones (that appear more in specialist academic discourse). This is an interesting point from both research and instruction points of view, showing that all of these MA students are in the development stage compared to expert academic writers. In 6.3.4, I also discussed this issue based on a rhetorical function (e.g., the revised CARS model as implemented in Lu et al., 2020) and demonstrated that both English L1 and ESL texts from the results sections produced larger numbers of finite dependent clauses that are associated with the step 'announcing and discussing results' based on the findings of Lu et al. (2020) on a large-scale study of expert/proficient writing.

Similarly, phrasal complexity (e.g., the use/amount of complex nominals, especially via CN/T), could distinguish between English L1 and EFL texts in most rhetorical sections. Mean length of T-unit and mean length of clauses were also among the measures that showed between-group differences in most rhetorical sections. Both of these syntactic constructs, however, only marked these differences with small-medium effects compared to much larger effects of subordination indices in distinguishing group differences. These findings are evidence that as the length of T-units or sentences increases, English L1 and ESL writers tend to use a greater number of dependent clauses compared to other structures (e.g., phrasal sophistication) in all sections of dissertations, while the EFL students use this space (the T-unit or sentence) to include more coordination. This latter conclusion can be readily observed in the dissertation excerpts in 6.3.4 where the EFL text has made extensive use of coordinate phrases compared to the other groups. The differences in the type of phrasal complexity features (complex nominals) in the English L1 vs EFL excerpts have been also discussed in 6.3.4. The English L1 and ESL students, for example, produced the pattern of ‘Noun + PP’ more frequently (nominal patterns) while the EFL text exhibited more verb phrases (verbal patterns). Biber and Gray (2010, p. 17) provide evidence on the “pervasive” use of “nominal/phrasal discourse style”, especially complex noun phrase constructions compared to verb phrases in proficient academic writing. They stress the pedagogical implications of such results in EAP courses, especially the understanding of the implicit nature of phrasal and nominal structures in proficient academic writing compared to a more explicit role of clausal embeddings, e.g., dependent clauses.

Finally, having the functional view in mind, I contextualised genre (e.g., academic writing in this study), task (MA dissertation), and rhetorical sections (the six sub-sections of theses/dissertations with different rhetorical functions) in the research design process and examined the effects of these rhetorical sections and students’ English language backgrounds on the values of the selected lexical and syntactic indices, then included them in the predictive classification models, and demonstrated the overall lexical and syntactic characteristics of each rhetorical section in the entire corpus. The results of mixed-effects models, for instance, show that the models with the interaction of rhetorical sections and groups of students can best explain the variations in most of the lexical and syntactic complexity indices used in this study, indicating that these two text-intrinsic and text-extrinsic variables are quite inter-dependent in accounting for the values of these complexity measures. In chapter four, I described at length the communicative functions of these rhetorical sections and the relationship between linguistic forms and the rhetorical functions of a text (see Lu et al., 2020 for example). This brings us to the importance of what Bhatia (1997b, pp. 143-147) addresses as genre-based ESP in academic settings, especially “familiarity with the dynamics of specialist genres, which includes the rhetorical forms and content” and “the awareness of the linguistic systems underlying a particular genre” by the learners whereby they could anticipate certain features of language as the realisation of specific genres and rhetorical sections. This is what Beers and Nagy (2009) refer to as the acquisition of genre-specific structures as the realisation of communicative goals of writing.

The functional view in this study also emphasises the role/function of English language backgrounds of the students and their academic contexts in the linguistic complexity differences of their dissertations. These findings are beneficial for L2/EFL teaching and learning, EAP writing instruction, and especially for materials developers and syllabus designers in EFL academic settings, e.g., advising them to explicitly address these concepts (e.g., the constructs of lexical density, diversity, sophistication) in postgraduate programmes, especially thesis and dissertation writing modules. As mentioned in chapter three, several studies such as Wolfe-Quintero et al. (1998) have acknowledged that the indices that represent/quantify such constructs can be used for testing and acquisition purposes. Gonzalez (2013), for instance, argues that lexical proficiency measures can prompt instructors to examine "if particular aspects of vocabulary knowledge are being overlooked in learner compositions" (p.15).

The differences between the EFL and ESL academic settings in this study were also discussed in detail in chapter five. I explained that the two groups have taken English proficiency tests before starting their MA programmes. The ESL students have had to pass the IELTS test with a minimum score of 7 and the EFL students have had to pass the MA entrance exam English proficiency test, a centralised high-stakes test, scored based on the norm-referenced system. Even though these tests differ, the entry to these courses is assumed to be based on an advanced-level proficiency in English. I also reviewed some aspects of academic writing and thesis writing modules in these settings. Acknowledging the fact that there could be different factors involved for the complexity differences of the dissertations, it seems plausible at this stage to consider the role of EAP academic immersion programmes more seriously with regards to the comparable (and in some cases, indistinguishable) performances of the ESL and English L1 students in the production of lexically and syntactically complex texts. This could be partly due to the shared syllabi, materials, and courses and the fact that the dissertations of both groups are examined with the same/similar criteria. The underlying assumption according to Ortega (2000) is that "L2 competence may proceed more slowly and might develop less fully in foreign language contexts than in second language contexts" (p. 72) and that EFL and ESL contexts constitute "distinct L2 populations" (p.512).

Previous scholars have also examined the contribution of academic immersion programmes on the linguistic proficiency of ESL students (for detailed discussions, see Bulté & Housen, 2014; Hinkel, 2004, as well as Mazgutova & Kormos, 2015). Bulté and Housen (2014), for instance, traced the development of lexical and syntactic proficiency of ESL students in an intensive academic writing programme and found that the values of nearly all syntactic complexity indices investigated in their study, as well as the subjective ratings of writing quality have significantly increased. Mazgutova and Kormos (2015) also emphasise that EAP programmes should explicitly address various lexical features of academic genres.

In her synthesis of studies on ESL and EAP academic writing programmes, Hinkel (2004) also shows that most studies unequivocally demonstrated that the knowledge of grammatical and syntactic structures and vocabulary plays a major role in academic writing success. This issue is further discussed in Swales (1990) and Biber (2006) in which 'the academic discourse community' expectations include strict forms of discourse construction as well as vocabulary and grammar use in academic registers.

The effectiveness of EAP academic immersion programmes has already been discussed in more detail in section 3.6. The few relevant studies that have investigated various lexical and syntactic complexity features are mainly developmental studies whereby the development of certain linguistic complexity features are traced for the same learners in their general English language production. This study, on the other hand, examines the possible effect of these programmes from a synchronous perspective, includes the comparison of other groups with different English language backgrounds, analyses a specialised academic writing corpus, and considers the effect of various rhetorical sections of the texts in linguistic complexity (differences) of their production. The results of this study, therefore, offers additional insight in this regard. However, due to mainly practical reasons, this study focused more on the quantitative aspects rather than a thorough inspection of such EAP programmes in different academic settings. This is because, unlike Iran, the UK universities do not follow a centralised education system and the syllabi and contents of these EAP programmes, especially academic writing modules (e.g., thesis/dissertation writing modules) differ.

Based on these considerations, the findings of this study and the differences of academic texts of EFL vs. ESL students, therefore, could be considered in EAP programmes: as most research in this area has mainly focused on the undergraduate EAP programmes, the EAP practitioners in the EFL settings could design short and intensive postgraduate EAP immersion programmes and/or thesis and dissertation writing immersion programmes to bridge this gap and/or to advance the linguistic proficiency of these students. Along this line, and emphasising the role of students' awareness, Silva (1993) proposes that students need to supply themselves with "a syntactic and lexical repertoire with which to produce more sophisticated academic texts" (p. 671). Similarly, Shahriari et al. (2017) believe that an awareness of such important linguistic features and structures by expert writers is crucial for the linguistic development of Iranian and other English L2 master's students.

7.4. Final Remarks: Limitations, Delimitations, Suggestions and Directions for Future Research

This in-depth investigation yielded substantial findings regarding the lexical and syntactic proficiency differences of English L1, EFL, and ESL postgraduate academic writers that, it is proposed, are useful for L2 teaching and learning, corpus-based research on discipline-specific linguistic features, academic (writing) immersion programmes, EAP writing pedagogy, and especially for materials developers and syllabus designers in EFL academic settings regarding thesis/dissertation writing modules, as well as for measure selection, evaluation, and testing processes, the use and/or development/modification of NLP tools in investigating linguistic complexity indices and/or features, and statistical modelling processes. Even though care has been taken to follow precise methodological and analytical principles and guidelines in conducting this research, I acknowledge that there are a number of limitations that could be addressed in future studies to further our understanding of these research results.

One such limitation is the size of the corpus that can affect the results and accuracy of models. In'nami and Koizumi (2011) for instance argued that sample size affects the results of model fit indices. Although Kline (2005) believes that a sample of 200 observations is large enough for statistical modelling, other researchers argue that the sample size is relative to the model complexity and the number of free parameters needed to be estimated (see for example Raykov & Marcoulides, 2006 and their discussions on a sample ten times larger than the model's free parameters). Apart from the number of parameters, In'nami and Koizumi (2011) emphasise that other determinants of a reasonable sample size are the strength of the relationship among the indicators, the type of indicators (categorical vs continuous), the type of estimator (e.g., robust maximum likelihood, etc) and the reliability of indicators. Since there is no consensus on the required sample size for statistical modelling and a reliable detection of linguistic patterns in a corpus, it is wise that future researchers collect larger amounts of texts for similar corpus-based studies, or as In'nami and Koizumi (ibid.) indicate, use the Monte Carlo analysis to find the required sample size. Due to copyright and/or the inaccessibility issues, collecting MA dissertations in various sub-disciplines of applied linguistics and language studies turned out to be a challenging task. I urge corpus designers and master's students to take the initiative to build a large-scale corpus of MA dissertations across disciplines and proficiency levels to aid future researchers in better capturing the linguistic characteristics of English L1 vs. L2 texts at the disciplinary and sub-disciplinary levels.

The type of the corpus, e.g., the genre, the number of rhetorical sections investigated, and the type and number of various disciplines or sub-disciplines also directly affect the results (e.g., the type of lexical and syntactic patterns observed and/or the results of linguistic predictors of classification models based on rhetorical sections), and hence the implications. Future projects could include various other academic writing genres (e.g., textbooks, journal articles, critical reviews, etc) and sub-genres (e.g., by using PhD theses alongside the MA dissertations) across various disciplines in the research design.

The inclusion of other theoretically-driven factors beside groups and rhetorical sections in the research design could also enhance the accuracy of classification models (e.g., the machine learning models) and influence the type of strong predictors of lexical and syntactic proficiency across classes (i.e., across groups and rhetorical sections in this study). Two examples could be the effect of gender and students' L1s as will be discussed in the following paragraphs. In keeping with this recommendation, experimental studies (e.g., via pedagogic interventions and explicit teaching of linguistic complexity indices) could systematically investigate the effect of ESL academic immersion programmes on the postgraduates' lexical and syntactic proficiency and performance differences, and on the quality of English academic writing. I also acknowledge the possible effect of proofreading (e.g., by students' supervisors) on lexical and syntactic choices; students, however, reassured me that this effect was minimal, with these reviewes mainly targeting the accuracy and spelling.

Some areas were also necessarily delimited due to practical reasons and/or because of the specific objectives of this study; addressing these areas could potentially reveal important findings as well.

I delimited the EFL group to Iranian MA students and the English L1 to British ones. An important reason was the validity of the results of the comparison of ESL and English L1 students that share the same academic context (i.e., UK universities); had I expanded the English L1 group to consider, for instance, students from the US, Australia, Canada, etc, it would have been methodologically less-accurate to discuss the results (e.g., similar performances of the ESL and English L1s) regarding the role of the ESL immersion academic programmes in the context of the UK. Restricting the EFL group to the Iranian ones was mainly due to accessibility reasons and for more specific targeted results for materials development in that context.

In this study, I also decided not to include the L1s of ESL and EFL students mainly because of the contradictory findings in this regard in the literature which did not provide sufficient empirical justification for including the L1s as other independent variables. I also had to avoid an overly-complex research design because of the already-complex and multi-layered nature of this research. Jarvis (2002) for example did not find any straightforward relationship between the L1 background of students and the lexical diversity values in their narrative writings; he believes that “this issue is far from being settled” (p. 75). Chen, Alexopoulou, and Tsimpli (2019), on the other hand, found evidence of the effect of students’ L1s (using 10 typologically diverse L1s in a general SLA corpus) on the values of some syntactic complexity structures, e.g., subordination structures, which they attribute to the syntactic differences between students’ L1 and English. Since this line of research in academic writing has not reached maturity yet, I call for further investigation in this area, e.g., by conducting a research similar to the present study and incorporating the first language of students in the research design to gauge the effect of L1s on proficiency differences, the values of measures, and the main characteristics of rhetorical sections of dissertations. This might show to be a key determinant of linguistic proficiency differences of English L1 vs. L2 academic writing and/or to play an important role in predictive classification models of linguistic proficiency.

The number and type of lexical and syntactic complexity indices were also limited to the 33 measures investigated in this study. These indices were carefully selected over several stages of measure evaluation to include indices that had previously been reported as effective/ strong indicators and/or predictors of linguistic proficiency in advanced learner and/or academic corpora of different types. I also included some similarly-computed indices with contradictory findings (or because of an absence of conclusive evidence) to test the effectiveness of each index in each pair/group of related indices, as a measure-selection guide for future researchers working on similar projects as this study. However, given the scope, time limit, and other practicalities of this project, I was not able to expand these sets of measures to all possible and/or effective indices. I, therefore, invite other researchers to continue this process of measure evaluation, selection, and testing with larger sets of indices.

Writing proficiency as one aspect of of linguistic proficiency, should not be restricted to the analysis of these lexical and syntactic measures either. Future studies could also consider other dimensions and features of proficiency such as lexical bundles and collocations, idiomatic use of vocabulary, lexical/syntactic/grammatical errors, lexical fluency, lexical networks, lexical density of word classes, as well as the relationship between various lexical and syntactic complexity measures for a holistic picture of linguistic proficiency and development of academic writers. In keeping with these recommended features and indices, researchers could also integrate qualitative and quantitative methods to leverage other potentials of corpus linguistics and for a more comprehensive inquiry into the effects of these indices and features on linguistic proficiency and/or performance (differences).

A thorough comparison of the most prominent lexical and syntactic characteristics of academic writing and academic spoken discourse is also another line of research which I suggest to future scholars. Despite some similarities in linguistic patterns across these two modes of academic language, their characteristics are yet to act in tandem. It would be interesting to examine the extent to which similar patterns would emerge in corpora of postgraduate students' speaking proficiency tests, lectures/conference talks, and other types of academic spoken discourse, using the same set of lexical and syntactic complexity measures, i.e., to discover the relationship between the two production modes of academic writing and academic spoken discourse and to verify the results of this study and the studies of researchers like Halliday (1985) and Lu (2012).

Finally, the investigation of the multidimensional constructs of lexical and syntactic complexity could be facilitated using various statistical modelling methods, especially various machine learning (ML) methods and their respective fast-growing algorithms with new and scientifically-driven capabilities and flexibilities that are yet to be explored in linguistics research. The potency of many ML methods to bypass the traditional assumptions associated with the conventional statistics (e.g., the assumptions of data and/or residuals' distributions, linear vs. non-linear issues, sample sizes, homogeneity of variance, (multi)collinearity, etc), make them among the most effective statistical methods for predictive models and classification systems, (see for instance the discussions in Norouzian, de Miranda, & Plonsky, 2018; Shalev-Shwartz, 2014; Probst et al., 2019; and Ziegler & König, 2014 among others). One such strong capability according to Shalev-Shwartz (2014, p. 6) is the efficiency in processing huge databases and detecting meaningful patterns "that are outside the scope of human perception". This, I believe, is particularly useful for corpus linguistics which is also primarily concerned with detecting linguistic patterns in large-scale naturally-occurring data in order to attain a holistic description of languages and language learning and development. Notable examples of research using various statistical modelling methods in corpus research and other linguistics research are Baumann and Winter (2018), Brown, et al. (2014), Gries (2019), Murakami et al. (2017), and Tagliamonte and Baayen (2012). Even though the advantages of using these methods outweigh the difficulties (see the discussions in 7.2.4), I caution future researchers about the interpretability issues of such methods, the selective use of model fit indices to report models' predictive powers, algorithm-specific issues (e.g., speed, accuracy, scalability, memory-intensiveness of some, etc), the optimisation issues (e.g., the sensitivity of some of the algorithms to hyperparameters and a lack of specific and agreed-upon methods in parameter tuning of some algorithms), and the classification difficulties in the presence of very similar categories (e.g., the ESL and English L1 groups in this study), just to name a few.

I find this research to be situated at the crossroads of corpus linguistics, EAP (especially academic writing research), SLA, quantitative linguistics (e.g., regarding the testing and operationalisation of linguistic indices), statistical modelling (and using programming languages like R in statistical analyses), computational linguistics (e.g., the use and modification of NLP tools) as well as theoretical and applied linguistic complexity. I hope this multidisciplinary project sets the direction for and inspires future works in advancing our understanding of the prominent characteristics of (postgraduate) academic writing and the effectiveness of various linguistic (complexity) indices in capturing English L1 vs. L2 texts using the principles of corpus linguistics and quantitative linguistics along with statistical modelling methods for building reliable predictive classification models. I further hope that these efforts ultimately translate into real-life practices, e.g., materials development (and by extension, syllabus design and pedagogical policies), revisiting the theories and/or assumptions regarding the nature of (English) L1 vs. L2 academic discourse for EAP and SLA practitioners, and corpus-driven instruction based on the realisation of genre, discipline-specific, and rhetorical functions of academic discourse. Finally I suggest the implementation of these real-life practices via corpus-based research on linguistic indices using statistical modelling methods and NLP tools by local practitioners in the EFL academic contexts as well as in the ESL academic immersion programmes to examine expected linguistic proficiency in these two settings.

Appendix A

A Discussion of Evidence of the Use of the IMRD Rhetorical Structure in Early Scientific Works: Cases of Ibn Al Haytham (a.k.a., Alhazen), Ptolemy, and Newton

Although an exact origin of the use of the IMRD rhetorical structure in the scientific works is not clear, two researchers of scientific history, Nader El-Bizri, a professor at the American University of Beirut, and Mark Smith, a professor at the University of Missouri, have identified a general structure of IMRD in several early scientific books as a response to my enquiry. The personal email communications with their exact wording appear in the following paragraphs.

Prof. Nader El-Bizri has identified an IMRD organisational pattern in three books of Ibn Al Haytham, books I-III of ‘Kitāb al-Manāzir’, translated as ‘The Book of Optics’ in the 11th century. Alhazen is referred to as the ‘father of modern optics’ and one of the early examples of conducting the ‘scientific method’ of experimentation (see the discussions in El-Bizri, 2005; Smith, 2004) and is renowned especially for the seven-volume book of optics. In a personal communication (January 2020), professor El-Bizri has confirmed the following about the IMRD rhetorical structure in each chapter in the first three volumes of this book:

> “Ibn al-Haytham proceeds by way of: {**I**} introducing the theme of his investigation (for example the nature of visual perception); he then notes {**M**} the existing theories or methods that are deployed in studying it (for example, explanations in terms of geometry [Euclid, Ptolemy] vs. physics [Aristotle]), and he proposes an outline of his own explications of the phenomenon being investigated and the method he will use (for example combining geometry with physics in mathematical modeling that informs and is guided by experimentation); then {**R**} he details his experimental work, his geometric models, his observations using direct visual perception or installations for controlled testing; then {**D**} he further elaborates on his findings and any drawbacks in the method or the observations and errors in data. However, this is not done across the seven books of his *Optics* (*Kitab al-manazir*/*De aspectibus* or *Perspectiva*) in this sequence, but we can, for instance, see that it resonates with his Books I-II and possibly III as well.”

Professor Mark Smith has also commented (in an email communication, January 2020) the following about the IMRD rhetorical structure in a letter written by Isaac Newton (this scientific letter is a modern equivalent of a scientific article that can be accessed via this URL http://www.newtonproject.ox.ac.uk/view/texts/normalized/NATP00006) as well as in the works of Ptolemy, a Greek mathematician in AD 170, and Ibn Haytham, an Iraqi scientist on Optics and mathematics in AD 965:

> "There are certainly examples of early scientific thinkers who followed the IMRD model virtually if not literally. Among them are Ptolemy and Ibn al-Haytham, who structured their optical analyses according to that model, albeit with an emphasis on **method** and **results**, but all in aid of demonstrating various universal characteristics of light and vision. A clearer case, perhaps, is Isaac Newton in his 1672 paper on light and colour, which is carefully constructed to highlight the **method** (i.e., instruments and procedures) as a means of **validating the results**, after which he was able to **conclude** that 'white' light is actually a composite of all radiant colors."

For further readings and an analysis of 'The Book of Optics' consult these references that are not cited in the References section of my thesis.

El-Bizri, N. (2005). A philosophical perspective on Alhazen's Optics. Arabic Sciences and Philosophy 15, 189-218. Cambridge: Cambridge University Press. doi:10.1017/S0957423905000172

Smith, M. (2004). What is the history of medieval optics really about? Proceedings of the American Philosophical Society, 148 (2), 180-194.

Appendix B
Complementary Results

A brief version of R code as well as the non-significant results (anova, mean differences, Tukey HSD, etc) for the lexical and syntactic measures and comparisons, and the VIF of fixed effects in the mixed-effect models A on both lexical and syntactic datasets are available at https://github.com/Maryam-Nasseri/lex-syn-modelling

Random forest second and third models on lexical and syntactic datasets with two groups only

In section 6.7.1, two sets of random forest models were conducted for predicting strong lexical and syntactic predictors of group membership for the three groups of postgraduate students (EFL, ESL, and English L1). However, as it is discussed in detail in section 6.7.2 and due to very similar performances of ESL and English L1 groups based on ANOVA tests, the first sets of random forest analysis could not efficiently classify each students' lexical and syntactic values to the correct group it belonged. As a result of the inclusion of these two similar groups, the accuracy of the models were lowered down as demonstrated in table 6.39. Therefore, second random forest analyses were conducted with EFL and English L1 groups only to test whether the models' accuracy improves as a result of excluding the ESL group. The secondary RF analysis on the lexical dataset with all lexical variables for EFL and English L1 groups resulted in more than 20% increase in lexical model's accuracy for classifying the groups and the similar analysis on the syntactic dataset with all syntactic variables resulted in more than 10% increase in accuracy. To rule out the possible effect of the number of classes in the increase in accuracy of the second models, third models were also constructed using the two similar groups of ESL and English L1 this time. The third models on both lexical and syntactic datasets obtained considerably lower accuracy as the second models, indicating that the number of classes (three vs. two) is not the primary cause of accuracy levels. The results of these second and third models for correctly classifying the relevant groups of students with other model performance indices are presented in below tables.

Random forest second models for predicting group membership with EFL and English L1 groups

Model	Accuracy [CI]	Precision	Recall (Sensitivity)	Specificity (True Negative Rate)	F1
Groups ~ 22 lex	78% [63- 90]%	74%	78%	79%	76%
Groups ~ 11 syn	57% [41- 72]%	50%	61%	54%	55%

Random forest third models for predicting group membership with ESL and English L1 groups

Model	Accuracy [CI]	Precision	Recall (Sensitivity)	Specificity (True Negative Rate)	F1
Groups ~ 22 lex	50% [34, 66]%	44%	67%	37%	53%
Groups ~ 11 syn	47% [32, 63]%	40%	44%	50%	42%

Appendix C

Appendix C1.

An explanation and the formulas of the performance indices of structural factor analyses discussed in section 6.5.

The following descriptions, explanations, and interpretation guides of models' fit/performance indices are based on Innami and Koizumi (2011), Hooper et al. (2008), Hu and Bentler (1999), Kline (2005), Lai and Green (2016), as well as MacCallum, Browne and Sugawara (1996).

--**Minimum Function Test Statistic** is the **Chi-Square** or χ^2_M which tests the model-implied against the observed variance/covariance matrix. This test is sensitive to collinearity: it produces larger values for larger correlation values among observed variables. (This model fit index examines an implied-model against the observed variance-covariance matrix). A non-significant chi-square is an indicator of a good model (it is sometimes called a 'badness-of-fit' index for this reason!). It is accompanied by the degrees of freedom and p-values to show if the chi-square values are statistically significant. This index is also sensitive to sample size.

-**RMSEA** index is the Root Mean Square Error of Approximation; it measures if the model can closely reproduce the data patterns. Values smaller than .08 are usually taken to indicate good model fit (ideally less than .06; zero indices the perfect fit). This index is usually accompanied by the 90% confidence intervals. RMSEA is sensitive to the size and distribution of the sample and the number of variables in the model; this might be due to including chi-square in the calculation process of RMSEA.

-**SRMR** or the Standardised Root Mean Square Residual is a residual-based fit index. SRMR transforms the predicted and sample covariance matrices to correlation matrices, so the values are the differences between the observed and predicted correlations. Values smaller than .08 is considered a good fit (another example of 'badness-of-fit label). In an ideal model, the residuals are close to zero.

-**Tucker-Lewis Index (TLI)** is a conservative comparative fit index which compares the fit of this study's model to the fit of the null model. A null model does not assume any covariances among the variables. Higher values of TLI, ideally closer to .9 represent a good fit or a good model. This index is sometimes called NNFI or the Non-normed Fit Index in the mentioned works.

There are other comparative fit indices such as CFI or the Comparative Fit Index, NFI or the Normed Fit Index, and other absolute fit indices such as GFI or the Goodness-of-Fit Index and its adjusted alternative, AGFI.

Appendix C2.

An explanation and the formulas of the performance indices of random forest models mentioned in section 6.7.

The performance indices in section 6.7 are derived from the confusion matrices of their corresponding models. A confusion matrix is a matrix of N x N cases, where N is the number of groups/cases in a model (e.g., a model with three groups of EFL, ESL, and NS has a 3 x 3 confusion matrix). The values in such a matrix represent the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) as explained below.

--True positives are all instances/data points where the algorithm correctly classified the case, e.g., correctly predicted the group membership.

--True negatives are all instances/data points/values of lexical and syntactic measures that the algorithm correctly decides where a data point does not belong to a group i.e., correct rejections or correct classification of negative cases.

--False positives or Type I error are all instances/values that the algorithm identified as positive (belonging to a particular class/group) but were actually negative (i.e., did not belong to that class/group).

--False negatives or Type II error are all instances that the algorithm identified as negatives (i.e., the algorithm predicted that value not to belong to a particular group) but were actually positive (i.e., that value actually belonged to that particular group/class).

Model Performance Indices:

These indices are derived from the confusion matrix of each model:

-**Accuracy** is the overall instances that the algorithm classified correctly (i.e., both true positives and true negatives).

Formula: TP + TN /total instances in a confusion matrix

--**Precision** or the **positive predictive value** is the ability of a model to identify ONLY the relevant data points or the correct classifications. It is an index of the exactness of a classifier. A large value of precision means there are less false positives/type I error.

formula: TP/TP+FP

--**Sensitivity** or **recall** (also known as true positive rate) is the ability of a model to find ALL relevant cases/correct classifications. A low recall value indicates many instances of type II error or false negatives and a large value of recall means less false negatives (i.e., less misdiagnosis).

formulas:TP/TP+FN

--**Specificity** or the **true negative rate** is the correct classification of negative cases.

formula: TN/TN + FP

--**F score** or the F1 is the weighted average of the recall and precision. This is a harmonic mean rather than a simple mean, because it punushes the extreme values and gives equal weights to both measures. There is a trade-off between the recall and precision, e.g., as we increase the precision, the recall is decreased; therefore, an ideal F1 score is closer to 1 which is an optimal balance of precision and recall. A large F score means less overall misdiagnosis/ misclassification and higher accuracy.

Formula: F1= 2 * precision*recall/ precision+ recall

– **Variable importance and its computation method**: Variable importance shows the importance of each predictor in the output of a random forest classification model. The value for each predictor is obtained via a graphical plot called ROC (Receiver Characteristic Curve) as a function of TPR (true positive rate) and FPR (false positive rate) on each axis. For a two-class model, the graph shows the performance at all classification thresholds usually at intervals/probabilities between 0 and 1. The computation finds the optimal threshold corresponding to better classification (detecting more true positives and minimising false positives). This threshold can be adjusted by the user based on what is important in a research design (e.g., getting more true positives, etc). The area under this optimal classification curve, is called AUC (Area Under the Curve) that shows the measure of performance of a model for all possible classification thresholds (i.e., how much the model can distinguish between the classes). Higher AUC curves are associated with better models and as the measure of variable importance. For a multi-class classification model, this process is repeated for all pairwise problems. This measure is scale-invariant and based on the ranking of the predictors rather than their absolute values.

Appendix D

How to Access and Use Lexical Complexity Analyzer for Academic Writing (LCA-AW)

LCA-AW (version 2.1; Nasseri & Lu, 2019), for Python 2 users, is a modification and an extension of Lexical Complexity Analyzer (LCA, 2012) developed by Xiaofei Lu, a professor of applied linguistics and Asian studies at The Pennsylvania State University. The original LCA analyser and its details can be found at this URL: https://aihaiyang.com/software/lca/. Both LCA and LCA-AW analyse 25 measures of lexical density, diversity/variation, and sophistication as elaborated in chapters two, five, and six in this thesis. Detailed explanation and discussions on LCA-AW are presented in chapters five (section 5.3.2) and seven (section 7.2.3). To calculate lexical sophistication, LCA-AW filters words through the British National Corpus (BNC) frequently-used word list (with an option to use the American National Corpus or ANC frequently-used word list) as well as the British Academic Written English (BAWE) word list for linguistics- and language-related disciplines.

Both analysers are open-source and free programmes that were developed in Python. LCA-AW code can be accessed and downloaded via this URL: https://www.researchgate.net/publication/330313752_LCA-AW_Lexical_Complexity_Analyzer_for_Academic_Writing

The README.txt file in the downloaded folder describes this analyser in more details and explains the required process for analysing texts. The texts to be analysed should be saved as plain texts with the .txt extension, there should not be any space nor special characters in the filenames, and all filenames should start with lowercase letters. Every text should have a minimum of 50 words. The texts then should be POS-tagged and lemmatised. The part-of-speech tagging can be done using the Stanford POS Tagger or Tree Tagger. Lemmatisation can be done using Morpha or Tree Tagger. Both LCA and LCA-AW run via Unix-based systems, e.g., Mac OS or Linux distributions. In the respective command line or terminal change the directory to the LCA-AW folder and run the following for analysing a single text:

```
python lc.py input_file > output_file
```

And the following for analysing a folder of texts:

```
python folder-lc.py path_to_folder > output_file
```

Replace the `input_file` with your input text filename and replace the `path_to_folder` with the full path to the folder that contains your texts or a single text to be analysed. The above examples use the BNC word list as the first word list to filter lexically sophisticated words. To use the ANC, replace the above script with `folder-lc-anc.py`. All scripts in the LCA-AW use the BAWE word list as well as the BNC or ANC word lists. The output file is a comma-delimited file with the lexical complexity labels and their respective values. For more information about processing files via the command line or terminal see Lu (2014).

References

Abdollahzadeh, E. (2011). Hedging in postgraduate student theses: A cross-cultural corpus study. *International Conference of Languages, Literature, and Linguistics,* 581-586. Singapore : IACSIT Pres,.

Adel, S. M. R., & Ghrobani Moghadam, R. (2015). A comparison of moves in conclusion sections of research articles in pyschology, persian literature and applied linguistics. *Teaching English Language*, 9 (2), 167-191.

Ai, H. & Lu. X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students writing. In N. Ballier, A. Díaz-Negrillo, and P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* , pp 249-264. Amsterdam: John Benjamins.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning,* 67, 180-208.

Allison, D., Cooley, L., Lewkowicz, J., & Nunan, D. (1998). Dissertation writing in action: The development of a dissertation writing support program for ESL graduate research students, *English for Specific Purposes,* 17 (2), 199-217. https://doi.org/10.1016/S0889-4906(97)00011-2

Amnuai, W. (2017). The textual organization of the discussion sections of accounting research articles, *Kasetsart Journal of Social Sciences*, 1-6.

Amnuai, W., & Wannaruk, A. (2013). Investigating move structure of English applied linguistics research article discussions published in international and Thai journals. *English Language Teaching*, 6(2), 1e13.

Anthony, L. (2018). AntConc, computer software, version 3.5.7. Tokyo, Japan: Waseda University.

Arnaud, P. J. L. (1984). The lexical richness of L2 written production and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, and D.K. Stevenson, (Eds.), *Practice and Problems in Language Testing, Papers from the International Symposium on Language Testing,* (pp. 14-28). Colchester: University of Essex.

Arthur, B. (1979). Short-term changes in EFL composition skills. Paper presented at *the 13th TESOL Convention* , Boston.

Baayen, R .H. (1989). *A corpus-based approach to morphological productivity*. *Statistical analysis and psycholinguistic interpretation*. Diss. Amsterdam: Free University.

Baayen, R. H. (2001). *Word frequency distributions.* Dordrecht: Kluwer.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. Philadelphia: University of Pennsylvania.

Banks, D. (2008). *The Development of Scientific Writing: Linguistic Features and Historical Context*. London: Equinox.

Bardel, C. & Gudmundson, A. (2012). Aspects of lexical sophistication in advanced learners' oral production: Vocabulary acquisition and use in L2 French and Italian. *Studies in Second Language Acquisition*, 34, 269-290.

Bardovi-Harlig, K. (1992). A second look at T-uni 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. *SSLA* , 11, 17-34.

Barker, M., & Pederson, E. (2008). Syntactic complexity and coordination in a verbal production task: Preliminary summary of experiment results. In T. Givón, and M. Shibatani (Eds.), *Syntactic complexity: Diachrony, acquisition, neuro-cognition, evolution* (pp. 391-404). Amsterdam : John Benjamins.

Barker, M., & Pederson, E. (2009). Syntactic complexity versus concatenation in a verbal production task. In T. Givón and Shibatani, M. (Eds.), *Syntactic Complexity: Diachrony, acquisition, neuro-cognition, evolution*, (pp.391-403). John Benjamins.

Barth, D., & Kapatsinski, V. M. (2018). Evaluating Logistic Mixed-Effects Models of Corpus-Linguistic Data in Light of Lexical Diffusion, In D. Speelman, K. Heylen, and D. Geeraerts (Eds.), *Mixed-Effects Regression Models in Linguistics.* Springer-Verlag, DOI: 10.1007/978-3-319-69830-4_6

Bartoń, K. (2019). MuMIn package in R, version 1.43.6. Available from https://cran.r-project.org/web/packages/MuMIn/index.html

Bar-Yam, Y. (2002). General features of complex systems. In *Encyclopedia of Life Support Systems (EOLSS), Knowledge Management, Organizational Intelligence and Learning, and Complexity*, vol. 1.

Bastardas-Boada, A. (2017). Complexity and Language Contact: A Socio-Cognitive Framework. In S. Mufwene, C. Coupé, & F. Pellegrino (Eds.), *Complexity in Language: Developmental and Evolutionary Perspectives* (Cambridge Approaches to Language Contact, pp. 218-244). Cambridge: Cambridge University Press. doi:10.1017/9781107294264.009

Basturkmen, H. (2009). Commenting on results in published research articles and masters dissertations in language teaching. *Journal of English for Academic Purposes*, 8(4), 241-251.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, 70, 20-38.

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

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

Behnam, B., & Mirzapour, F. (2012). A comparative study of intensity markers in engineering and applied linguistics. *English Language Teaching*, 5 (7), 158-163. doi:10.5539/elt.v5n7p158

Berman, R. A. (2009). Developing linguistic knowledge and language use across adolescence. In E. Hoff & M. Shatz (Eds.), *Blackwell handbook of language development.* (pp. 347-367). Malden, MA : Wiley-Blackwell.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Bhatia, V. K. (1997a). Genre-mixing in academic introductions. English for Specific Purposes, 16, 181–195.

Bhatia, V. K. (1997b). Applied genre analysis and ESP. In T. Miller (ed.), *Functional approaches to written text: Clssroom applications*. English Language Programs: United States Information Agency.

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

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

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

Biber, D., & Gray, B. (2013). Identifying multi-dimensional patterns of variation across registers. In M. Krug and J. Schlüter (eds.), *Research Methods in Language Variation and Change*, pp. 402-420.Cambridge: Cambridge University Press.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

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

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5), 639-669.

Bitchener, J. (2010). W*riting an applied linguistics thesis or dissertation: A guide to presenting empirical research*. Basingstoke: Palgrave MacMillan.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis*, 20(1), 1-49.

Bloch, J. (2008). *Technology in the second language composition classroom*. Ann Arbor: University of Michigan Press.

Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effect. *Elementary Education*, 49, 513-516.

Boulesteix, A., Janitza S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining Knowl Discov*, 2, 493-507.

Brett, P. (1994). A genre analysis of the results section of sociology articles. *English for Specific Purposes*, 13 (1), 47-59

Brown, C., T. Snodgrass, M. A. Covington, R. Herman & S. J. Kemper (2007). Measuring propositional idea density through part-of-speech tagging. Poster presented at *Linguistic Society of America Annual Meeting*, Anaheim, California.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London Lund Corpus of English Conversation. *Behavior Research Methods, Instrumentation & Computers*, 16, 502–532. doi:10.3758/BF03200836

Brown, L., Winter, B., Idemaru, K., & Grawunder, S. (2014). Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*, 66, 45-60.

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. T*he Language Teacher*, 4 (37), 13-16.

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock. G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance, *Aphasiology,* 14, 71-91.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42-65.

Bunton, D. (1998). *Linguistic and textual problems in PhD and M.Phil. theses: an analysis of genre moves and metatext*. Unpublished PhD Thesis, University of Hong Kong.

Bunton, D. (1999). The use of higher level metatext in PhD theses. *English for Specific Purposes*, 18 (Supplement 1), S41–S56.

Bunton, D. (2002). Generic moves in PhD thesis introductions. In J. Flowerdew (Ed.), *Academic discourse* (pp. 57–75). London: Pearson Education.

Bunton, D. (2005). The structure of PhD Conclusion chapters, *Journal of English for Academic Purposes*, 4, 207-224.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Canty, A., & Ripley. B. (2017). Boot: Bootstrap Functions. R package version 1.3-20. Retrieved from https://CRAN.R-project.org/package=boot

Cappelli, G. (2010). Lexical complexity: theoretical and empirical aspects. In L. Pinnavaia and N. Brownlees (Eds.), *Insights into English and Germanic lexicology and lexicography: past and resent perspectives, (pp.*115-127). Monza/Italy: Polimetrica International Scientific Publisher.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing* 3, 179–201.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12, 267-296.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43th Annual Meeting of the ACL*, 173-180.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition*, 12, 43–64.

Chen, X., Alexopoulou, T., & Tsimpli, I. (2019). L1 effects on the development of L2 subordination. Poster presented at *the 29th conference of the European Second Language Association (EuroSLA 29)*. Lund, Aug. 28-31.

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, (pp. 113-119). Osaka, Japan.

Chen, M. & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. Pr*oceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, (pp. 722-731), Portland: Oregon.

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13, 53-76.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1971). Deep structure, surface structures, and semantic interpretation. In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics*. New York: Cambridge University Press.

Chung, S., Chao, F. Y. A., & Hsieh, Y. (2009). VocabAnalyzer: A referred word list analysing tool with keyword, concordancing and N-gram functions. *23rd Pacific Asia Conference on Language, Information and Computation*, 638-645.

Ciani, A. J. (1976). Syntactic maturity and vocabulary diversity in the oral language of first, second, and third grade students. *Research in the Teaching of English*, 10, 150–56.

Cilliers, P. (1998). *Complexity and Postmodernism: Understanding Complex Systems*. Routledge, London.

Cobb, T. (2019). Web Vocabprofile. Retrieved from http://www.lextutor.ca/vp

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Complexity (2010). In A. Stevenson (Ed.), *Oxford Dictionary of English*. Oxford: Oxford University Press.

Complexity (2019). *Online Etymology Dictionary*, retrieved from https://www.etymonline.com/search?q=complexity

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

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research* , 69 (5), 176-183.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment*, 10(7), 1-9.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.

Coxhead, A. J. (1998). *An academic word list* (English Language Institute Occasional Publication, No. 18). Wellington, New Zealand: Victoria University of Wellington.

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

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, 45, 355–362. doi:10.5054/tq.2011.254528

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. Routledge.

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches*. California, USA: SAGE Publications.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(4), 367–83.

Crossley, S.A., Allen, D., & McNamara, D.S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16 (1), 89-108.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count based and band based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41, 965–981. doi:10.1016/j.system.2013.08.002

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment*, 7 (1), article 74.

Crossley, S.A., & McNamara. D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18, 119-135.

Crossley, S. A., & McNamara, D. S. (2010). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-135.

Crossley, S. A., & McNamara, D. S. (2012a). Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In S. Jarvis, & S. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 106–126). Bristol: Multilingual Matters.

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, 17, 171–192.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66-79.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307-334.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing*, 28, 561–580.

Crowhurst, M. (1983). Syntactic complexity and writing quality: A review. *Canadian Journal of Education*, 8, 1–16.

Crystal, D. (1982). *Profiling linguistic disability*. London: Edward Arnold.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modelling and longitudinal data analysis, In Plonsky (Ed.), 159-180, *Advancing Quantitative Methods in Second Language Research*, New York: Routledge.

Cutler, A. (1983). Lexical complexity and sentence processing. In G. B. Flores d'Arcais and R. J. Jarvella (Eds.), *The process of language understanding*, (pp43-79). University of California: J. Wiley.

Dahl, O. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In D.H. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledg*e (pp. 234-244). Cambridge: Cambridge University Press. Chapter DOI: http://dx.doi.org/10.1017/CBO9780511667268.016

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24, 197-222.

Daller, H., & Xue, H. (2009). Vocabulary knowledge and academic success: A study of Chinese students in UK higher education. In: Richards B., Daller M.H., Malvern D.D., Meara P., Milton J., Treffers-Daller J. (Eds.), *Vocabulary Studies in First and Second Language Acquisition*, (pp. 179-193). London: Palgrave Macmillan.

Dascălu, M., Trausan-Matu, S., & Dessus, P. (2012). Towards an integrated approach for evaluating textual complexity for learning purposes. In E. Popescu, Q. Li, R. Klamma, H. Leung, & M. Specht (Eds.), *Advances in web-based learning*, ICWL 2012, Volume 7558 of the series, *Lecture Notes in Computer Science*, (pp 268-278). Springer-Verlag: Berlin Heidelberg.

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

Day, R. (1989). The origins of the scientific paper: The IMRAD format, *The American Medical Writers Association Journal*, 4 (2), 16-18.

DeBoer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System*, 47, 139-145.

Derakhshan, A., & Karimian Shirejini, R. (2020). An Investigation of the Iranian EFL Learners' Perceptions Towards the Most Common Writing Problems. *SAGE Open*. https://doi.org/10.1177/2158244020919523

Deutscher, G. (2009). Overall complexity: A wild goose chase? In G. Sampson, Gil, D., and Trudgill, P. (Eds.), *Language Complexity as an Evolving Variable*, (pp. 243–251). Oxford: Oxford University Press.

Dewaele, J.M., & Pavlenko, A. (2003). Productivity and lexical diversity in native and non-native speech: A study of cross-cultural effects. In Cook, V. (Ed.), *Effects of the second language on the first*, UK: Multilingual Matters, Ltd.

Dickinson, D. K. (2001). Large-group and free-play times: Conversational settings supporting language and literacy development. In D. K. Dickinson and P. O. Tabors (Eds.), *Beginning literacy with language: Young children learning at home and at school* (pp. 223- 255). Baltimore, MD: Paul h. Brooks Publishing.

Dickinson, D. K., & Tabors, P. O. (Eds). (2001). *Beginning literacy with language: Young children learning at home and at school.* Baltimore, md: Paul H. Brookes Publishing.

Di Domenico, E. (2017). Introduction, In E. Di Domenico (Ed.), *Syntactic complexity from a language acquisition perspective*, (pp. 1-22). Cambridge Scholars Publishing.

Dobakhti, L. (2016). A genre analysis of discussion sections of qualitative research articles in applied linguistics. *Theory and Practice in Language Studies*, 6 (7), 1383-1389. DOI: http://dx.doi.org/10.17507/tpls.0607.08

Dontcheva-Navratilova, O. (2018). Intercultural and interdisciplinary variation in the use of epistemic lexical verbs in linguistics and economics research articles. *Linguistica Pragensia*, 28 (2), 154-167.

Doró, K. (2008). *The Written Assessment of the Vocabulary Knowledge and Use of English Majors in Hungary* (Doctoral Dissertation). Graduate School in Linguistics: University of Szeged.

Doró, K. (2015). Changes in the lexical measures of Undergraduate EFL students' argumentative essays. In P. Pietilä, K. Doró, and R. Pípalová (Eds.): *Lexical issues in L2 writing*, (pp. 57-76). Cambridge Scholars Publishing.

Drożdż, S., Kwapień, J., & Orczyk, A. (2009). Approaching the linguistic complexity. In Zhou J. (Eds.), *proceedings of the International Conference on Complex Sciences, Complex Sciences*. [Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 4.] (pp. 1044-1050). Berlin, Heidelberg :Springer. DOI: https://doi.org/10.1007/978-3-642-02466-5_104

DuBay, W.H. (2004). *The Principles of Readabilit*y. Available at: http://www.impact-information.com/impactinfo/readability02.pdf.

Dudley-Evans, T. (1986). Genre analysis: An investigation of the introduction and discussion sections of MSc. dissertations. In M. Coulthard (Ed.), *Talking about text* (Discourse Monograph No. 13, pp.128- 145.). English Language Research, University of Birmingham.

Dudley-Evans, T. (1994). Genre analysis: an approach to text analysis for ESP. In M. Coulthard (Ed.), *Advances in Written Text Analysis*, (pp. 219–228). London: Routledge.

Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théoretique du vocabulaire? *Le francais moderne*, 46, 25-32.

Durán, P., Malvern, D., Richards, B. & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, 25 (2), 220-242.

Durrant, P. (2014). Discipline- and level-specificity in university students written vocabulary. *Applied Linguistics*, 35(3), 328-356.

Edwards, J. (1994). *Multilingualism*. London: Penguin.

Egbert, J. and Plonsky, L. (2015). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*, 10 (3), 291-313.

Ellis, R. (2003). Task-based language learning and teaching. Oxford: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30(4), 474-509.

Ellis, R. & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *SSLA* , 26, 59-84.

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

Esmaeili F., & Esmaeili, K. (2015). Academic writing experience of Iranian postgraduate students. *Anglisticum Journal (IJLLIS)*, 4 (5), 8-13.

Evert, S., Wankerl, S., & Nöth, E. (2017). Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. *In the Proceedings of Corpus Linguistics International Conference*, University of Birmingham, Birmingham, UK, July.

Farahani, A. A. K., & Meraji, S. R. (2011). Cognitive Task Complexity and L2 Narrative Writing Performance. *Journal of Language Teaching and Research*, *2*(2), 445-456.

Farvardin, M. T., Afghari, A., & Koosha, M. (2012). Analysis of four-word lexical bundles in Physics research articles. *Advances in Digital Multimedia (ADMM)*, 1(3), 134-139.

Feinerer I, Hornik K (2018). tm: Text Mining Package. R package version 0.7-5, https://CRAN.R-project.org/package=tm.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, (pp. 277-298). John Benjamins. https://doi.org/10.1075/lllt.32

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

Flahive, E. D., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oiler, Jr., and K. Perkins (Eds.), *Research in language testing*, (pp.171-176). Rowley, Massachusetts: Newbury House.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233.

Flowerdew, J. (2007). The non-Anglophone scholar on the periphery of scholarly publication. *AILA Reviews*, 20, 14-27. doi 10.1075/aila.20.04flo

Flowedew, J. (2015). Some thoughts on English for Research Publication Purposes (ERPP) and related issues. *Language Teaching*, 48 (2), 250-262. doi:10.1017/S0261444812000523

Flowerdew, J. (2017). Corpus-based approaches to language description for specialized academic writing. *Language Teaching*, *50*(1), 90-106. doi:10.1017/S0261444814000378

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language*. New York: McGraw-Hill.

Foster, P & Skehan, P. (1999). The influence of source planning and focus of planning on task-based performance. *Language Teaching Research* , 3 (3), 215-247.

Foster, P. & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning* , 59 (4), 866-896.

Fox, J. (2014). NumSummary: Rcmdr package. R package version 2.0-4. Available from https://www.rdocumentation.org/packages/Rcmdr/versions/2.0-4/topics/numSummary

Fradis, A., Mihailescu, L., & Jipescu, I. (1992). The distribution of major grammatical classes in the vocabulary of Romanian aphasic patients. *Aphasiology*, 6(5), 477-489.

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written English*. Princeton, NJ: Educational Testing Service.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computation, and theoretical perspectives* (pp. 129-189). New York: Cambridge University Press.

Fries, C. C. (1952). *The Structure of English*. New York: Harcourt Brace.

Friginal, E. (2013). Developing research report writing skills using corpora. *English for Specific Purposes*, 32, 208–220.

Friginal, E., Li, M., & Weigle, S. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, 23, 1-16.

Garrard, P., Maloney, L. M., Hodges, J. R., & Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, 128(2), 250–260.

Gerlanc, D., & Kirby. K. (2013). BootES: Bootstrap effect sizes. R package version 1.01. Available from https://CRAN.R-project.org/package=bootES

Ghavamnia, M., Tavakoli, M., & Esteki, M. (2013). The effect of pre-task and online planning conditions on complexity, accuracy and fluency on EFL learners written production. *Porta Linguarum*, 20, 31-43.

Gholami, J., Mosalli, Z., & Bidel Nikou, S. (2012). Lexical complexity and discourse markers in soft and hard science articles. *World Applied Sciences*, 17(3), 368-374.

Gillaerts, P., & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes*, 9, 128-139.

Goldfield, B. (1993). Noun bias in maternal speech to one-year-olds. *Journal of Child Language*, 20, 85-99.

Golebiowski, Z. (2009). Prominent messages in education and applied linguistics abstracts: How do authors appeal to their respective readers? *Journal of Pragmatics*, 41, 753-769.

Golub, L. S., & Fredrick, W. C. (1971). Linguistic structures in the discourse of fourth and sixth graders. *Report from project on Reading and Related Language Arts*, Center for Cognitive Learning, Madison, Wisconsin.

Gonzalez, M. (2013). *The Intricate Relationship between Measures of Vocabulary Size and Lexical Diversity as Evidenced in non-Native and Native Speaker Academic Compositions* (Doctoral Dissertation). College of Education and Human Performance: University of Central Florida.

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

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123-145.

Green, C. (2019). A multilevel description of textbook linguistic complexity across disciplines: Leveraging NLP to support disciplinary literacy. *Linguistics and Educations*, 53, 1-11. https://doi.org/10.1016/j.linged.2019.100748

Gregori-Signes, C., & Clavel-Arroitia. B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia- Social and Behavioral Sciences*, 198, 546-556.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13, 403–437.

Gries, S. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models, *Corpora*, 10 (1), 95-125. DOI: 10.3366/cor.2015.0068

Gries, S. (2019). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, aop, De Gruyter, 1-31. https://doi.org/10.1515/cllt-2018-0078

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

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. [The statistical characteristics of vocabulary: An essay in metholdogy.] Paris: Presses Universitaires de France.

Halekoh U, Højsgaard S (2017). pbkrtest: Parametric Bootstrap and Kenward Roger Based Methods for Mixed Model Comparison. R package version 0.4-7. https://CRAN.R-project.org/package=pbkrtest.

Halleck, G. B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal*, 79 (2), 223-234.

Halliday, M.A.K. (1985). *Spoken and Written Language*. Geelong, Vic.: Deakin University Press.

Halliday, M. A. K. (1987). Spoken and written modes of meaning. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 55-72). New York: Academic Press.

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

Halliday, 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* (pp. 241–277). Norwood, NJ: Ablex.

Harfitt, G. J. (1999). *A comparison of lexical richness in samples of written and spoken English from a group of secondary six students in Hong Kong*. Master's Dissertation, Hong Kong University Theses Online (HKUTO).

Harley, B., & King. M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, 11, 415–439.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal*, 80 (3), 309-326.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.

Hershberger, S., Marcoulides, G., & Parramore, M. (2003). Structural equation modeling: An introduction. In B. Pugesek, A. Tomer, & A. Von Eye (Eds.), *Structural Equation*

Modeling: Applications in Ecological and Evolutionary Biology, (pp. 3-41). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511542138.002

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

Hinkel, E. (2004). *Teaching Academic ESL Writing: Practical Techniques in Vocabulary and Grammar*. Routledge.

Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan*, 2, 21-30.

Holmes, R. (1997). Genre analysis,and the social sciences: an investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes*, 16(4), 321-337.

Holmes, D. I., & Singh, S. (1996). A stylometric analysis of conversational speech of aphasic patients. *Literary and Linguistic Computing*, 11, 133-140.

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

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 11, 45-60.

Hopkins, A., & Dudley-Evans, T. (1988). A Genre-Based Investigation of the Discussion Sections in Articles and Dissertations. *English for Specific Purposes*, 7, 112-121.

Housen, A., Bulté, B., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time – the case of Dutch-speaking learners of French in Brussels. *French Language Studies*, 18, 1-22.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, 35(1), 3-21.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency - Definitions, measurement and research. In A. Housen, V. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA*, (pp. 1-20). Amsterdam: John Benjamins.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55, DOI: 10.1080/10705519909540118

Hu, G., & Cao, F. (2011). Hedging and boosting in abstracts of applied linguistics articles: A comparative study of English- and Chinese-medium journals. *Journal of Pragmatics*, 43, 2795-2809. doi:10.1016/j.pragma.2011.04.007

Hu, Z., Brown, D., & Brown, L. (1982). Some linguistic differences in the written English of Chinese and Australian students. *Language Learning and Communication*, 1 (1), 39-49.

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

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Urbana, IL: The National Council of Teachers of English.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first?. *TESOL Quarterly,* 4 (3), 195-202.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics*, 23 (2), 215-239.

Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing*, 13, 133-151.

Hyland, K. (2008). Genre and academic writing in the disciplines, *Plenary Speeches, a revised version of a plenary paper prsented at the bianual conference of the European Association of the Teaching of Academic Writing,* 30 June 2007, Bochum, Germany. doi:10.1017/S0261444808005235

Hyland, K. (2009). English for professional academic purposes: Writing for scholarly publication. In D. Belcher (Ed.), *English for specific purposes in theory and practice* (pp. 83-105). Ann Arbor, MI: University of Michigan Press.

Hyland, K. (2015). Genre, discipline and identity. *Journal of English for Academic Purposes,* 19, 32-43. https://doi.org/10.1016/j.jeap.2015.02.005

Hyland, K. (2016). General and specific EAP. In K. Hyland, and P. Shaw (Eds.), *The routledge handbook of English for academic purposes,* (pp. 17-29). Oxon: Routledge.

Hyland, K. (2017). Learning to write for academic purposes: Specificity and second language writing. In Bitchenet, J., Storch, N. & Witte, R. (Eds.). *Teaching Writing for Academic Purposes to Multilingual Students: Instructional Approaches*. London: Routledge.

Hyland, K., & Jiang, F. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes*, 46, 1-14. https://doi.org/10.1016/j.esp.2016.11.001.

Hyland, K., & Shaw, P. (2016). *The Routledge handbook of English for academic purposes*. Oxon: Routledge.

Hyland, K., & Tse, P. (2005). Hooking the reader: a corpus study of evaluative that in abstracts. *English for Specific Purposes*, 24, 123-139. https://doi.org/10.1016/j.esp.2004.02.002

Hyland, K., & Tse, P. (2007). Is there an academic vocabulary? TESOL Quarterly, 41(2), 235-253.

In’nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly,* 8, 250-276. doi:10.1080/15434303.2011.565844

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4, 51–69.

Ishikawa, T. (2006). The effect of task complexity and language proficiency on task-based language performance. *The Journal of Asia TEFL,* 3 (4), 193- 225.

Jalali, H., & Ghayoomi, S. (2010). A comparative qualitative study of lexical bundles in three academic genres of Applied Linguistics. *MJAL*, 2 (4), 323-333.

Jalilifar, A., & Dabbi, R. (2012). Citation in Applied Linguistics: Analysis of introduction sections of Iranian master’s theses. *Linguistik Online 57,* 7 (12), 91-104.

Jalilifar, A., Firuzmand, S., & Roshani, S. (2011). Genre analysis of problem statement sections of MA proposals and theses in Applied Linguistics. *The International Journal - Language Society and Culture*, 33, 88-93.

Jalilifar, A., & Vahid Dastjerdi, H. (2010). A contrastive generic analysis of thesis and dissertation abstracts: Variations across disciplines and cultures. *Journal of Persian Language and Literature*, 26, 20-48.

James, K. (1984). The writing of theses by speakers of English as a foreign language: the results of a case study. In R. Williams, J. Swales and J. Kirkman (Eds.), *Common Ground: Shared Interests in ESP and Communication Studies*, (pp. 99-113). ELT Documents 117. Oxford: Pergamon Press & The British Council.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57–84.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63, 87-106.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, 34(4), 537-553.

Jarvis, S. (2018). Personal communication on the effectiveness of the MATTR measure of lexical diversity for finding academic proficiency differences.

Jarvis, S., & Crossly, S. A. (2012). *Approaching language transfer through text classification: Explorations in the detection-based approach*. Bristol: Multilingual Matters.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, 12, 377-403.

Jiang, F., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves, *English for Specific Purposes*, 46, 1-14.

Johansson, J. & Geisler, C. (2011). Syntactic aspects of the writing of Swedish L2 learners of English. *Language and Computers*, 73(1),139–55.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics*, 53, 61-79.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, 37, 13-38.

Johnson, R. L. (1979). Measures of vocabulary diversity. In D. E. Ager, F. E. Knowles, & J. Smith (Eds.): *Advances in Computer-Aided Literary and Linguistic Research, (pp. 213–227)*. University of Aston: Department of Modern Languages.

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, 56, 1-15.

Johnson, M. D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. Jo*urnal of Second Language Writing*, 21, 264-282.

Joseph, R., Lim, J., & Nor, N. (2014). Communicative moves in forestry research introductions: Implications for the design of learning materials. *Procedia Social and Behavioral Sciences*, 134, 53-69.

Joye, S. (2004). ESL students and formal accuracy. *English Leadership Quarterly*, 27 (1), 13-16.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31-36.

Kameen, P. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins & J. Schachter (Eds.), *On TESOL '79: The Learner in Focus, (pp. 343-364)*. Washington, D.C.: TESOL.

Khany, R., & Khosravian, F. (2013). The investigation of RAS in Applied Linguistics: Convergence and divergence in Iranian ELT context. *Proceeding of the Global Summit on Education*, (pp.143-150). 11-12 March, Kuala Lumpur.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24, 269-292. doi:10.1016/j.esp.2004.08.003

Karimnia, A. (2013). Writing research articles in English: Insights from Iranian university teachers of TEFL. *Procedia - Social and Behavioral Sciences*, 70, 901 – 914.

Kim, J.Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, 69 (4), 27-50.

Klee., T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*, 12, 28-41.

Klein, E. (1966). *A comprehensive etymological dictionary of the English language*. Amsterdam : ELSEVIER Publishing Company.

Klein, D., & Manning, C. (2003). Fast exact inference with a factored model for natural language parsing. In A*dvances in Neural Information Processing Systems 15 (NIPS)*, (pp. 3-10). Cambridge, MA: MIT Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling, 2nd* edition. New York: The Guilford Press.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1), 60- 69. doi: http://dx.doi.org/10.7820/vli.v01.1.koizumi

Kol, S., & Schcolnik, M. (2008). Asynchronous formus in EAP: Assessment issues. L*anguage Learning and Technology*, 12 (2), 49-70.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, 20, 148-161.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26.

Kuiken, F., & Vedder, I. (2008a). Task complexity, task characteristics and measures of linguistic performance. In S. van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 113-125). (Contactforum). Brussel: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

Kuiken, F., & Vedder, I. (2008b). The influence of task complexity on linguistic performance in L2 writing and speaking: The effect of mode. C*onference Proceedings in 32th LAUD SYMPOSIUM*, March 10-13, Germany.

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in l2 writing and speaking. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 143–170). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian, and Spanish. *International Journal of Applied Linguistics*, 29, 192-210.

Kuiken, F., Vedder, I. & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, I. Vedder & G. Pallotti (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research (EUROSLA monographs series, 1)* (pp. 81-99). [S.l.]: European Second Language Association.

Kwan, B. S. C. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes*, 25 (1), 30-55.

Kyle, K. (2011). *Objective measures of writing quality*. Unpublished MA dissertation. Colorado State University.

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. PhD Dissertation, Georgia State University. https://scholarworks.gsu.edu/alesl_diss/35

Kyle. K. (2018). TAALED: Tool for the automatic analysis of lexical diversity (beta version 1.2.4, Python-based). https://www.linguisticanalysistools.org/taaled.html

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757-786.

Kyle, K., & Crossley, S. **(**2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34,12-24.

Kyle, K., & Crossley, S. **(**2017). Assesing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513-535.

Kyle, K., Crossley, S., & Kim, Y. (2015). Native language identification and writing proficiency. *International Journal of Learner Corpus Research*, 1 (2), 187-209.

Laflair, G. T., Egbert, J., & Plonsky. L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests and ANOVAs. In L. Plonsky. (Ed.): *Advancing quantitative methods in second language research*. Oxon: Routledge.

Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree, *Multivariate Behavioral Research*, 51(3), 220-239, DOI: 10.1080/00273171.2015.1134306

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL QUARTERLY*, 12 (4), 439-448.

Larsen-Freeman, D. (1983). Assessing global second language proficiency, In H. Seliger and M. Long (Eds.), *Classroom-oriented Research in Second Language Acquisition*. Newbury House.

Larsen-Freeman, D. (1997). Chaos complexity science and second language acquisition. Applied Linguistics, 18(2), 141-165.

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. doi:10.1093/applin/aml029

Larson-Hall, J. (2016). *A guide to Doing Statistics in Second Language Research Using SPSS and R*. (2nd ed.). Oxon: Routledge.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, 25(2), 21–33. doi:10.1177/003368829402500202

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16 (3), 307-322.

Lee, L. (1974). *Developmental sentence analysis*. Evanston, IL: Northwestern University Press.

Leki, I. (1991). *Understanding ESL writers.* Portsmouth, NH: Boynton / Cook.

Levshina, N. (2015). *How to do linguistics with R. Data exploration and statistical analysis.* John Benjamins.

Levy, R. & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, (pp. 2231–2234)*. Genoa, Italy: ELRA.

Lewkowicz, J. & Cooley, L. (1995). *Postgraduate Students' Writing Needs and Difficulties*. Hong Kong: The English Centre, The University of Hong Kong.

Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System*, 28, 229-245.

Lim, J. M. H. (2006). Method sections of management research articles: A pedagogically motivated qualitative study. *English for Specific Purposes*, 25, 3, 282-309.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer, (Eds.), *L2 vocabulary knowledge, acquisition and use: New perspectives on assessment and corpus analysis, Eurosla Monographs Series, 2,* (pp.109-126).

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *International Review of Applied Linguistics in Teaching*, 49 (3), 221- 240.

Linnarud, M. (1975). Lexis in free production: An analysis of the lexical texture of Swedish students' written work. *University of Lund, Department of English: Swedish-English Contrastive Studies*, report number 6.

Linnarud, M. (1983). On lexis: The Swedish learner and the native speaker compared. In K. Sajavaara (Ed.), *Cross language analysis and second language acquistion* (pp. 249-261). Jyvaskyla: University of Jyvaskyla.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund, Sweden: CWK Gleerup.

Liu, L., & Li., L. (2016). Noun phrase complexity in EFL academic writing: A corpus-based study of postgraduate academic writing. *The Journal of Asia TEFL*, 13 (1), 48-65.

Loban, W. (1976). *Language development: Kindergarten through grade twelve.* Research Report No. 18. Urbana, IL: National Council of Teachers of English.

Long, S. H., Fey, M. E. & Channell, R. W. (2008). *Computerized Profiling, v*ersion 9.7.0. Cleveland, OH: Case Western Reserve University.

Lorés, R. (2004). On RA abstracts: from rhetorical structure to thematic organization. *English for Specific Purposes*, 23, 2, 280-302.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14, 3–28.

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, 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96 (2), 190-208.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 16-27. http://dx.doi.org/10.1016/j.jslw.2015.06.003

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34(4), 493-511.

Lu, X., J. E. Casal and Y. Liu. 2020. 'The rhetorical functions of syntactically complex sentences in social science research article introductions', *Journal of English for Academic Purposes*, 44: 1-16.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum Associates.

Maleki, A., & Zangani, E. (2007). A survey on the relationship between English language proficiency and the academic achievement of Iranian EFL students. *Asian EFL Journal*, 9(1), 86-96.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language*, (pp. 58–71). Clevedon, UK: Multilingual Matters.

Malvern, D. D., & Richards, B. J. (2000). Validation of a new measure of lexical diversity. In M. Beers, B. van den Bogaerde, G. Bol, J. de Jong, and C. Rooijmans (Eds.), *From sound to sentence: Studies on first language acquisition*. Groningen: Centre for Language and Cognition.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave MacMillan.

Mancilla, R. L., Polat, N., & Akcay, A. O. (2015). An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions. *Applied Linguistics*, 1-24. doi:10.1093/applin/amv012

Manning, C., Grow, T., Grenager, T., Finkel, J, & Bauer, J. (accessed 2018). PTBTokenizer, from Stanford Tokenizer, accessed from https://nlp.stanford.edu/software/tokenizer.shtml

Mao, Z., & Jiang, L. (2017). Exploring the effects of the continuation task on syntactic complexity in second language writing. *English Language Teaching*, 10(8), 100-106.

Marcus, M., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2), 313–330.

Mazgutova, D., & Kormos. J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing*, 29, 3-15.

McArdle, J. J. (2011). Some ethical issues in factor analysis. In A. T. Panter & S. K. Sterba (Eds.), *Multivariate applications series. Handbook of ethics in quantitative methodology*, (pp. 313-339). New York, NY, US: Routledge.

McCarthy, P. (2005). A*n assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Unpublished PhD dissertation, University of Memphis.

McCarthy, P. (2020). Personal communication on the effectiveness of the two measures of vocd-D and HD-D for capturing lexical diversity of long texts based on sample sizes and their underlying formulas.

McCarthy, P. , & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488.

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42 (2), 381-392. doi:10.3758/BRM.42.2.381

McCarthy, P.M., Watanabe, S., & Lamkin, T.A. (2012). The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. In P.M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution*, (pp. 312-333). Hershey, PA: IGI Global.

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*, 2, 141–154.

McCarthy, P., Watanabe, S., & Lamkin, T. A. (2012). The Gramulator: A tool to identify differential linguistic features of correlative text types. In P. McCarthy, and C. Boonthum-Denecke, (Eds.), *Applied Natural Language Processing: Identification, Investigation and Resolution*. DOI: 10.4018/978-1-60960-741-8.ch018.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

McKee, G, Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literacy and Linguistic Computing*, 15, 323-337.

McNamara, D.S., Crossley, S.A., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication*, 27 (1), 57-86. DOI: 10.1177/0741088309351547

McWhorter, J. (2001). The world’s simplest grammars are creole grammars. *Linguistic Typology*, 6, 125–166.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge, England: Cambridge University Press.

Meara, P. (2005a). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, S. Christopher, M. Á. Gómez González, & S. M. Doval Suárez (Eds.), *The dynamics of language use*, (pp. 271–285). Amsterdam: John Benjamins.

Meara, P. (2005b). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, 26 (1), 32-47.doi: 10.1093/applin/amh037

Meara, P., Lightbown, P. M., & Halter, R. (1997). Classrooms as lexical environments, *Language Teaching Research*, 1, 28–47.

Ménard, N. (1983). *Mesure de la richesse lexicale*. Geneva: Slatkine.

Mendelsohn, D. J. (1981). We should assess lexical richness, not only lexical error. P*roceedings of TESOL convention*, Detroit.

Michéa, R. (1971). De la relation entre le nombre des mots d'une fréquence déterminée et celui des mots différents employés dans le texte. *Cahiers de Lexicologie*, 18, 65–78.

Michel, M., Murakami, A., Alexopoulou T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, 3(2), 124-152.

Miller, J. F. (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research in child language disorders: A decade of progress,* (pp. 211–220). Austin, TX: Pro-Ed.

Miller, J. F. (1996). Progress in assessing, describing, and defining child language disorder. In K. N. Cole, P. S. Dale, and D. J. Thal (Eds.), *Assessment of communication and language,* (pp. 309-324). Baltimore, MD: Paul Brooks Publishing.

Milton, J. & Tsang, E. S. C. (1993). A corpus-based study of logical connectors in EFL students' writing: Directions for future research. In R. Pemberton & E.S.C. Tsnag (Eds.), *Lexis in studies,* (pp. 215- 246). Hong Kong: Hong Kong University Press.

Minnen, G., Carroll, J. and Pearce, D. (2001). Applied morphological processing of english. *Natural Language Engineering*, 7, 207–223.

Moiinvaziri, M. (2012). Analyzing vocabulary used in Payame Noor University general English textbook: A Corpus Linguistic approach. *Frontiers of Language Teaching*, 3, 204- 212.

Monroe, J. H. (1975). Measuring and enhancing syntactic fluency in French. *The French Review*, 48 (6), 1023-1031.

Moritz, M. W., Meurer, J. L., & Dellagnelo, A. (2008). Conclusions as components of research articles across Portuguese as a native language, English as a native language and English as a foreign language: A contrastive genre study the. *Specialist*, 29(2), 233-253.

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of TESL student performance. *System*, 32, 75–87. doi:10.1016/j.system.2003.05.001

Muangsamai, P. (2018). Analysis of moves, rhetorical patterns and linguistic features in New Scientist articles. *Kasetsart Journal of Social Sciences*, 39, 236-243.

Mufwene, S. S., Coupé, C., & Pellegrino, F. (2017). *Complexity in language: Developmental and evolutionary perspectives*. Cambrdige: Cambridge University Press.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus, *Corpora,* 12 (2), 243-277.

Nakagawa, S., & Schielzeth, H. (2013) A general and simple method for obtaining R² from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution*, 4, 133–142.

Nasseri, M. (2017). A Corpus-based Analysis of Syntactic Complexity measures in the Academic Writing of EFL, ESL, and Native English Master's Students. *The 9th International Corpus Linguistics Conference*, University of Birmingham, 25-28 July 2017. http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper37.pdf

Nasseri, M., & Lu, X. (2019). Lexical Complexity Analyzer for Academic Writing (LCA-AW), software, version 2.1. DOI:10.5281/zenodo.2537862.

Nation, P. (1984). *Vocabulary lists: Words, affixes and stems*. Wellington, New Zealand: Victoria University of Wellington English Language Institute. Occasional Publication No. 12.

Nation, P. (1990). *Teaching and learning vocabulary*. New York: Newbury.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139524759

Nation, P. (2013). Learning vocabulary in another language. 2nd edition. Cambridge: Cmbridge University Press.

Nation, P., & Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Naves, T., Torras, M. R., & Celaya, M. L. (2003). Long term effects of an earlier start: An analysis of EFL written production. *EUROSLA Yearbook 3*, (pp. 103-129).

Nayar, P. B. (1997). ESL/EFL dichotomy today: Language politics or pragmatics? *TESOL Quarterly*, 31(1), 9-37.

Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe and M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*, (pp. 53-65). Routledge.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008-2010). The British Academic Written English (BAWE) corpus.

Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. In: Geoffrey Sampson, David Gil, and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 64–79. Oxford: Oxford University Press.

Nihalani, N. K. (1981). The quest for the L2 index of development. *RELC Journal*, 12, 50-56.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis, *Language Learning*, 50, 417–528.

Norris, J. M. & Ortega, L. (2009). Toward on organic approach to investigating CAF in instructed SLA: The case of complexity, *Applied Linguistics*, 30(4), 555–78.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second langage research: An applied approach. *Language Learning*, 68(4), 1032-1075.

Noyau, C., & Paprocka, U. (2000). La représentation de structures événementielles par les apprenants: granularité et condensation. *Roczniki Humanistyczne* 48 (5), 87-121.

Nwogu, K. N. (1997). The medical research paper: Structure and function. *English for Specific Purposes*, 16(2), 119–138.

O'Dowd, E. (2012). The development of linguistic complexity: A functional continuum. *Language Teaching*, 45, 329 - 346. doi: 10.1017/S0261444810000510

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing*, 22, 545–565.

Olinghouse, N. G., & Wilson. J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, 26, 45–65.

O’Loughlin, K. (1995). Lexical density in candidate output of direct and semi-direct versions of an oral proficiency test. *Language Testing,* 12, 217–237.

Onslow, M., Ratner, N. B., & Packman. A. (2001). Changes in linguistic variables during operant, laboratory control of stuttering in children, *Clinical Linguistics and Phonetics*, 15, 651-62.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners*. Ph.D. thesis, Manoa, HI: University of Hawaii.

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. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, 29*,* 82-94.

Ouellet, C., Cohen, H., Le Normand, M.-T., & Braun. C. (2000). Asynchronous language acquisition in developmental dysphasia, *Brain and Cognition*, 43, 352-357.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31 (1), 117-134.

Paltridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes*, 21, 125–143.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Park, S-Y. (2012). A corpus-based study of syntactic complexity measures as development indices of college-level L2 learners’ proficiency in writing. *Korean Journal of Applied Linguistics*, 28(3), 139-160.

Peacock, M. (2002). Communicative moves in the discussion section of research articles. *System*, 30, 479-497.

Peng, J. (1987). Organisational features in chemical engineering research articles. *ELR Journal* 1, 79-116.

Perfetti, C. A. (1985). *Reading ability*. Oxford: Oxford University Press.

Peristeri, E., Andreou, M., & Tsimpli, I. M. (2017). Syntactic and Story Structure Complexity in the Narratives of High- and Low-Language Ability Children with Autism Spectrum Disorder. *Frontiers in psychology*, *8*, 1-16. doi:10.3389/fpsyg.2017.02027

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. T*ESOL Quarterly*, 14(1), 61-69.

Perkins, M. (1994). Repetitiveness in language discorders: A new analytical procedure. *Clinical Linguistics and Phonetics*, 8, 321-336.

Pho, P. D. (2008). Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. *Discourse Studies*, 10 (2), 231-250.

Pica, T. (1984). L1 transfer and L2 complexity as factors in syllabus design, *TESOL Quarterly*, 18 (4), 689-704.

Pica, T. (1985). Linguistic simplicity and learnability: Implications for language syllabus design. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition,* (pp. 137-153). San Diego, CA: College Hill Press.

Pienemann, M. (1985). Learnability and syllabus construction. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition*, (pp. 23–77). San Diego, CA: College Hill Press.

Pietilä, P. (2015). Lexical diversity in L2 academic writing: A look at M.A. thesis conclusions. In P. Pietilä, K. Doró, and R. Pípalová (Eds.), *Lexical issues in L2 writing, (pp. 105-125)*. Cambridge Scholars Publishing: Newcastle upon Tyne.

Plonsky, L. (2015). *Advancing Quantitative Methods in Second Language Research*. New York: Routledge.

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in Applied Linguistics: Assessing its potential using shared data. *Applied Linguistics*, 36(5), 591-610.

Plonsky, L., & Oswald. F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878-912.

Pojanapunya, P., & Todd, W. R. (2011). Relevance of findings in results to discussion sections in applied linguistics research. In *Proceedings of the International Conference Doing Research in Applied Linguistics* (pp. 51-60). King Mongkut's University of Technology Thonburi and Macquarie University, Thailand.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47 (1), 101-143.

Polio, C., & Yoon, H-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. International Journal of Applied Linguistics, 28(1), 1-24.

Probst, P, Wright, M., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs, Data Mining and Knowledge Discovery*, 9:e1301, https://doi.org/10.1002/widm.1301

Python Software Foundation. Python Language Reference, versions 3.6.3 and 3.7.1. Available at http://www.python.org

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rafiei, K., & Modirkhamene, S. (2012). Thematicity in published vs, unpublished Iranian TEFL theses. Th*eory and Practice in Language Studies*, 2 (6), 1206-1213. doi:10.4304/tpls.2.6.1206-1213

Rafoth, B. A. , & Combs, W. (1983). Syntactic complexity and reader's perception of an author's credibility. *Research in the Teaching of English*, 17(2), 165- 169.

Rahimivand, M., & Kuhi, D. (2014). An exploration of discoursal construction of identity in academic writing. *Procedia- Social and Behavioral Sciences*, (pp.1492-1501). doi: 10.1016/j.sbspro.2014.03.570

Rahimpour, M., & Safarie, M. (2011). The effects of on-line and pre-task planning on descriptive writing of Iranian EFL learners. *International Journal of English Linguistics*, 1 (2), 274- 280.

Ramires, V. (2017). Genres' Analysis in Academic Contexts: The Abstract. *American Journal of Linguistics*, 5(1): 15-21. DOI: 10.5923/j.linguistics.20170501.03

Rawson, K.A. (2004). Exploring automaticity in text processing: syntactic ambiguity as a test case. *Cognitive Psychology*, 49(4), 333-69.

Raykov, T., & Marcoulides, G. A. (2006). *A First Course in Structural Equation Modeling* (2nd ed.). Mahwah, NJ: Erlbaum.

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity and lexical ambiguity. *Memory & Cognition*, 14 (3), 191-201.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Englewood Cliffs, New Jersey: Prentice Hall.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in Language Assessment*, (pp. 41-60). Mahwah, NJ: Lawrence Erlbaum.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych psych package in R version 1.8.12.

Richards, B. J. (1990). *Language development and individual differences: A study of auxilliary verb learning*. Cambridge: Cambridge University Press.

Richards, B. J., and Malvern, D. D. (2004). Investigating the validity of a new measure of lexical diversity for root and inflected forms. In K. Trott, S. Dobbinson and P. Griffiths (Eds), *The child language reader* (pp. 81–9). London: Routledge.

Rinker, T. (2017). *qdap: Quantitative Discourse Analysis Package*. Version 2.3.0. http://github.com/trinker/qdap.

Rinker, T. (2018). textclean: Text Cleaning Tools, version 0.9.3. Buffalo, New York. https://github.com/trinker/textclean

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22, 27 – 57.

Robinson, P. (2005). Cognitive complexity and task sequencing: studies in a componential framework for second language task design. *International Review of Applied Linguistics*, 43, 1-32.

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(3), 161-176.

Rosenberg, S. & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8 (1), 19-32.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36.

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. *Journal of Second Language Writing*, 29, 51-63.

Sadeghi, K., & Mosalli, Z. (2012). The effect of task complexity on fluency and lexical complexity of EFL learners' argumentative writing. *International Journal of Applied Linguistics and English Literature*, 1(4), 53-65.

Sadeghi, K. & Shirzad Khajepasha, A. (2015) Thesis writing challenges for non-native MA students, *Research in Post-Compulsory Education*, 20 (3), 357-373.

Safnil, A. (2014). The discourse structure and linguistic features of research article abstracts in English by Indonesian academics. *The Asian ESP Journal*, 10 (2), 191-224.

Sahragard, R., Baharloo, A., and Soozandehfar, S. M. (2011). A Closer Look at the Relationship between Academic Achievement and Language Proficiency among Iranian EFL Students. *Theory and Practice in Language Studies*, 1 (12), 1740-1748.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *Englisg for Specific Purposes*, 11, 93-113.

Salah, G. (1990). Discourse analysis and embedding depth of utterances: Clause analysis technique as a measure of complexity. In L. A. Arena (Ed.), *Language proficiency: Defining, teaching, and testing*, (pp. 121-128). New York: Plenum.

Salimi, A., Dadashpour, S., & Asadollahfam, H. (2011). The effect of task complexity on EFL learners' written performance. *Procedia-Social and Behavioral Sciences*, *29*, 1390-1399.

Salmani Nodoushan, M. A., & Khakbaz, N. (2011). A structural move analysis of discussion sub-genre in Applied Linguistics. *International Journal of Language Studies*, 5(3), 111-132.

Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24, 141-156.

Sánchez, J. A. (2018). Applicability and variation of Swales' CARS model to applied linguistics article abstracts, *ELIA*, 18, 213-240.

Sato, C. J. (1990). *The syntax of conversation in interlanguage development*. Tübingen: Gunter Narr.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, 11 (1), 1-22.

Schcolnik, M. (2008). Asynchronous forums in EAP: Assessment issues. *Language Learning & Technology*, 12 (2), 49-70.

Schulz, S. (2012). Thinking in propositions - Propositional idea density as a cross-language complexity measure. Tech. rep., Eberhard Karls University of Tübingen.

Scott, M. (2020). WordSmith Tools version 8, Stroud: Lexical Analysis Software. https://lexically.net/publications/citing_wordsmith.htm

Shah, S. K., Gill, A. A., Mahmood, R., & Bilal. M. (2013). Lexical richness, a reliable measure of intermediate L2 learners' current status of acquisition of English language. J*ournal of Education and Practice,* 4(6), 42-47.

Shahriari, H., Ansarifar, A, & Pishghadam, R. (2017). Phrasal complexity in the writing of Iranian EFL college-level students. *Proceedings of the International Corpus Linguistics Conference CL2017*, Birmingham, UK.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge: Cambridge University Press.

Sharma, A. (1980). Syntactic maturity: Assessing writing proficiency in a second language. In R. Silverstein (Ed.), *Occasional Papers in Linguistics*, No. 6, (pp. 318-325). Carbondale, IL: Southern Illinois University.

Sheehan, K. M., Kostin, I. Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards.* ETS Research Report RR-10-28. Princeton, NJ: ETS.

Shi, H., & Wannaruk, A. (2014). Rhetorical structure of researcg articles in agricultural science, *English Language Teaching*, 7 (8), 1-13. http://dx.doi.org/10.5539/elt.v7n8p1

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology,* 10, 1–40.

Sichel, H. S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, 11, 45-72.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27 (4), 657-675.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers,* 4, 26-36.

Skehan, P. (1989). *Individual differences in second language learnin*g. London: Edward Arnold.

Skehan, P. (1992). Second language acquisition strategies and task-based learning. *Valley University Working Papers in English Language Teaching*, Volume 1, 178-208.

Skehan, P. (1996). Second language acquisition research and task-based instruction, In J Willis, and D. Willis (Eds.), *Challenge and change in language teaching*, (pp. 17-30). Oxford: Heinemann.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language learning tasks. In B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application,* (pp. 107–124). London: Palgrave Macmillan.

Skehan, P. (2009b). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30 (4), 510- 532.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1 (3), 185- 211.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: a meta-analysis of the Ealing research, In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, and I. Vedder (Eds.), *Complexity, Accuracy, and Fluency in Second Language Use, Learning, and Teaching*. University of Brussels Press.

Snow, C. E., Tabors, P. O., & Dickinson, D. K. (2001). Language development in the preschool years. In D. K. Dickinson and P. O. Tabors (Eds.). *Beginnign literacy with language: Young children learning at home and at school* (pp. 1-25). Baltimore, MD: Paul H. Brooks Publishing.

Somers, H. H. (1966). Statistical methods in literary analysis. In J. Leeds (Ed.), *The computer and literary style,* (pp. 128-140). Kent, OH: Kent State University Press.

Sparks, J. (1988). Syntactic complexity, error and the holistic evaluation of ESL student essays. *The ORTESOL Journal*, 9, 35-49.

Štajner, S., & Mitkov, R. (2011). Diachronic stylistic changes in British and American varieties of 20th century written English language. *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, (pp.78-85). Hissar: Bulgaria, 16 September.

Štajner, S., & Mitkov, R. (2012). Style of religious texts in 20th century. *Semantic Scholar*. Retrieved from https://www.semanticscholar.org/paper/Style-of-religious-texts-in-20th-century-Stajner-Mitkov/5664148bd7cd0a0532c2653c6cbe9e71eaf4615d

Štajner, S., & Zampieri, M. (2013). Stylistic changes for temporal text classification. In *Proceedings of the 16th International Conference on Text, Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence -LNAI, 8082. Springer*, (pp. 519-526).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26 (4), 471-495.

Staples, S., Egbert, J., Biber, D., & Gray. B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, 33 (2), 149-183.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist*, 38, 878–893.

Stewart, M. F., & Grobe, C. H. (1979). Syntactic maturity and mechanics of writing: Their relationship to teachers' quality ratings. *Research in the Teaching of English*, 13, 207–216.

Stockwell, G. & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO Journal*, 20 (2), 337-359.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323.

Stubbs, M. (1994). Grammar, text and ideology: computer-assisted methods in the linguistics of representation. *Applied Linguistics*, 15, 201-223.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press. DOI:10.1017/CBO9781139524827

Swales, J., & Feak, C. B. (1994). *Academic writing for graduate students: Essential tasks and skills*. University of Michigan Press: Ann Arbor.

Swales, J. M., & Najjar, H. (1987). The writing of research article introductions. Wr*itten Communication*, 4, 175–191.

Szmrecsanyi, B. & Kortmann, B. (2012). Introduction: Linguistics complexity- Second language acquisition, indigenization, contact. In Kortmann Bernd & Benedikt Szmrecsanyi (eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact, (pp. 6-34)*. Berlin: de Gruyter.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. 6th edition, Pearson Education Inc.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/ were variation as a case study for statistical practice. *Language variation and change*, 24(2), 135-178.

Thomas, D. (2005). Type-token ratios in one teacher's classroom talk: An investigation of Lexical complexity. (University of Birmingham Essay Bank). Last retrieved 25 December 2015 from http://www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/research/resources/essays/language-teaching.aspx

Thompson, P. (1999). Exploring the contexts of writing: Interviews with PhD supervisors. In P. Thompson (Ed.), *Issues in EAP writing research and instruction* (pp. 37-54). Reading, UK: Centre for Applied Language Studies, University of Reading.

Thompson, P. (2002). Modal verbs in academic writing, in B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Anakysis*, (pp. 305-325). Amsterdam: Rodopi. DOI:https://doi.org/10.1163/9789004334236_023

Thompson, P. (2012). Thesis and Dissertation Writing. In B. Paltridge and S. Starfield (Eds). *The Handbook of English for Specific Purposes,* (pp. 283-299), Wiley-Blackwell. doi:10.1002/9781118339855.ch15

Thompson, P. (2016). Genre approaches to theses and dissertations. In K. Hyland and P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes*, chapter 29. London: Routledge, https://doi.org/10.4324/9781315657455

Thompson, P., Hunston, S., Murakami, A. and Vajn, D. (2017). Multi-dimensional analysis, text constellations, and interdisciplinary discourse. *International Journal of Corpus Linguistics*, 22(2), pp.153-186. https://doi.org/10.1075/ijcl.22.2.01tho

Tobias, S., & Carlson, J. E. (2010). Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate Behavioral Research*. (the original article published in 1969, 4(3), 375-377.

Torruella, J., & Capsada, R. (2013). Lexical statistics and tipological structures: A measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95, 447-454.

Toutanova, K., Klein, D., Manning C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pp. 252-259.

Treffers-Daller, J. (2013) Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In Jarvis, S. and Daller, M. (eds.) (pp. 79-104) *Vocabulary knowledge: human ratings and automated measures. Benjamins, Amsterdam*.

Treffers-Daller, J., Parslow, P., & Williams. S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, amw009:1-27.

Tseng, F. (2011). Analyses of move structure and verb tense of research article abstracts in applied linguistics journals. *International Journal of English Linguistics*, 1(2), 27-39.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and Humanities*, 32, 323-352.

Ullman, J. B. (2013). Structural equation modelling, In B. G. Tabachnik and L. S. Fidell (Eds.), *Using multivariate statistics,* (pp. 681-785). Pearson Education Inc.

Ure, J. (1971). Lexical density and register differentiation. In G.E. Perren & J.L.M. Trimm (Eds.), *Applications of Linguistics: selected papers of the 2nd International Congress of Applied Linguists*, pp. 443-452. Cambridge: Cambridge University Press.

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

Vajjala, B. S. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Doctoral Thesis. Retrieved from https://publikationen.uni-tuebingen.de/xmlui/handle/10900/64359

van Gijsel, S., Speelman, D., & Geeraerts, D. (2006). Locating lexical richness: A corpus linguisttic, sociovariational analysis. *8 ^{es} Journées internationales d'Analyse statistique des Données Textuelles*, 953-963.

van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, and J. Treffers-Daller. (Eds.): *Modelling and Assessing Vocabulary Knowledge*, (pp. 93-115). Cambridge: Cambridge University Press.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17 (1), 65-83.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Boogards and B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing*, (pp. 173-189). Amsterdam: Benjamins.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21, 239-263.

Vidaković, I., & Barker, F. (2009). Lexical development across second language proficiency levels: A corpus-informed study. In *Proceedings of the BAAL Annual Conference, 143-146*, Newcastle University.

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, 11–30.

Waldvogel, D. A. (2014). An analysis of Spanish L2 lexical richness. A*cademic Exchange Quarterly*, 18 (2), 17-24.

Waskita, D. (2008). Differences in men's and women's ESL academic writing at the University of Melbourne. *Journal Sosioteknologi Edisi 14 Tahun, 7*, 448-463.

Weaver, W. (1948). Science and complexity. *American Scientist*, 36, 536-544.

Wei, T. and Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Weissberg, R., & Buker, S. (1990). *Writing up research: experimental research report writing for students of English.* Prentice Hall Regents.

West, G. (1980). That-nominal constructions in traditional rhetorical divisions of scientific research papers. TE*SOL Quarterly*, 14 (4), 483-488.

Wickham, H. (2018). stringr package in R, version 1.3.1. https://cran.r-project.org/web/packages/stringr/index.html

Wilcox, R. (2011). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press.

Wimmer, G., and Altmann, G. (1999). Review article: on vocabulary richness. *Journal of Quantitative Linguistics*, 6, 1–9.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. New York: Routledge, https://doi.org/10.4324/9781315165547

Witte, S. P. & Davis, A. S. (1982). The stability of T-unit length in the written discourse of college freshmen: A second study. *Research in the teaching of English*, 16 (1), 71-84.

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

Woodward-Kron, R. (2008). More than just jargon: The nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes*, 7(4), 234-249.

Wray, A. (2017). The language of dementia science and the science of dementia language: linguistic interpretations of an interdisciplinary research field. *Journal of Language and Social Psychology*, 36 (1) , 80-95. DOI: 10.1177/0261927X16663591

Wright, M. N., & Ziegler, A. (2017). ranger : A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17.

Yang, R., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, 4 (22), 365-384.

Yang, W., Lu, X., & Weigle, S. A. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67.

Yngve, V. (1960). A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104, 444-466.

Yoder, P. J., Davies, B., & Bishop, K. (1994). Adult interaction style effects on the language sampling and transcription process with children who have developmental disabilities. *American Journal on Mental Retardation*, 99, 270-282.

Yoneoka, D. and Ota, E. (2017). Evaluating association between linguistic characteristics of abstracts and risk of bias: Case of Japanese randomized controlled trials. *PLOSONE*, 12 (3), e0173526. https://doi.org/10.1371/journal.pone.0173526

Yoon, H. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66, 130-141.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus usein L2 writing. *Journal of Second Language Writing*, 13, 257-283.

Yoon, H.-J., & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*. doi: 10.1002/tesq.296

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zarei Chamani, N., Pazhakh, A., Jalilifar, A., & Gorjian, B. (2012). Towards the establishment of an academic world list (AWL) for the abstract section of research articles. *Advances in Asian Social Science (AASS)*, 2 (2), 456- 460.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency—variable sensitivity. *Studies in Second Language Acquisition*, 27, 567–595.

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real-world applications. *WIREs Data Mining Knowl Discov*, 4, 55-63.

Zipf. G. K. (1932). Selected studies of the principle of relative frequency in language. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1935). *The psycho-biology of language*. Oxford, England: Houghton, Mifflin.

Zuur, A.F., Ieno, E.N., & Elphick, C.S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14.