A CORPUS-BASED COMPARISON OF THE ACADEMIC ESSAY WRITING OF BRITISH AND HONG KONG STUDENTS

by

Andrew John Morrall

Presented to the Faculty of Humanities
The Hong Kong Polytechnic University in Partial Fulfillment of
the Requirements for the Degree of

DOCTOR OF APPLIED LANGUAGE SCIENCES

THE HONG KONG POLYTECHNIC UNIVERSITY

OCTOBER 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________________ (Signed)

Andrew John Morrall (Name of student)

Abstract

Abstract of thesis entitled: ‘*A corpus-based comparison of the academic essay writing of British and Hong Kong students’, submitted by Andrew John Morrall for the degree of Doctor of Applied Language Sciences at The Hong Kong Polytechnic University in November 2019.*

This thesis compares two corpora of academic writing, one by native English speakers and the other by Hong Kong learners of English, in order to analyse the differences in language use between them, and from this make recommendations regarding the content of academic English courses.

The conceptual framework is one of Contrastive Interlanguage Analysis (Granger, 1996) and the research methodology is corpus-based linguistics.

The literature on this topic shows concerns about the usefulness of university English courses (Evans and Morrison, 2012), and possible solutions suggested by Hyland (2008, 2015) and Gardner (2012) in the field of genre and discipline analysis. The literature also contains a number of conclusions from previous research on corpora of students’ academic writing, and this thesis examines whether these can be applied to an interlanguage analysis of Hong Kong student’s writing.

Two main corpora were analysed, the PolyU Learner English Corpus (PLEC) and the British Academic Written English (BAWE) corpus. In order to give a better comparison, academic essays by native English speakers in year one were extracted to form a sub-corpus called BAWE-EON, and this was used as a reference corpus.

The research questions were firstly, to what extent are the commonly-taught aspects of academic essay writing and findings from the research literature on academic writing reflected in differences between the corpora, and secondly what changes to teaching and learning would these differences suggest?

The findings indicate that corpus comparisons are often not generalisable to other corpora, due mainly to the context of the corpus collection resulting in different language use in the texts. Factors in this include writers' proficiency level, genre and disciplinary variation. Applying theories from previous corpus comparisons gave rise to a number of recommendations for areas in which Hong Kong university students could improve their academic writing.

Based on the contrastive analysis of the two corpora, recommendations are given for the content of academic English writing courses, including suggestions for indirect applications, in which corpus linguistics is used outside the classroom, for example in planning curricula and materials, and suggestions for direct applications of corpus linguistics in data-driven learning by students using software such as concordancers inside the classroom. The importance of the selection of corpora and examples of language use that are most suitable for the context of students is emphasised, and methods of applying corpus linguistics techniques and tools to the specific context of a class of students are explained.

Limitations of the research include that while the BAWE texts were written for disciplinary courses, the PLEC essays addressed general topics and were written for a timed English course assignment, which limits their comparability. In addition, it is not known to what extent the BAWE students had been trained in academic writing, whereas the PLEC students were receiving instruction in it, which may have affected their use of language features, and thus the frequency comparisons in this research.

Acknowledgements

I would like to thank my supervisor, Dr. Xu Xuenfeng, for the valuable advice and support he has given me in the writing of this thesis.

I would like to thank Dr Linda Lin not only for her kind permission to use the PolyU Learner English Corpus (PLEC) corpus in this study, and Profs. Martin Warren and Winnie Cheng for permission to use the CJA14 corpus, but also for their support.

My thanks also go to the compilers of the British Academic Written English (BAWE) corpus which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics, Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (School of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800). My thanks to all involved and for permission to use BAWE.

The Louvain Corpus of Native English Essays (LOCNESS) Corpus is used courtesy of the Centre for English Corpus Linguistics (CECL), Université catholique de Louvain, Belgium.

The Multidimensional Analysis Tagger uses the Stanford Tagger (Toutanova, Klein, Manning, and Singer 2003), and my gratitude goes to the Stanford NLP Group.

My gratitude also goes to friends and colleagues who gave me valuable comments on the draft of this thesis: Mr James Nicoll, for his TEFL-teacher's point of view,

Dr Jozsef Szakos for his perspective on it as a linguist of Chinese, and Dr K.P. Kwan for his very valuable suggestions regarding the statistics.

Finally, I would like to thank my colleagues past and present for their support and encouragement, and my wife and son for their tolerance of my absences and distractedness over the last few years.

Table of Contents

List of Illustrative Materials

Glossary of Terms and Abbreviations

Annotation
The process of encoding linguistic information such as the part of speech of a word in a corpus.

Authentic
Examples of language that were not written specifically for inclusion in a corpus.

AFL
Academic Formulas List: a list of phrases that are used in academic texts significantly frequently.

AWL
Academic Word List: a list of words that are used in academic texts significantly frequently

BAWE
British Academic Written English corpus: a corpus of merit- or distinction-graded academic texts from British universities.

BAWE-EON
A sub-corpus of BAWE used in this thesis, containing only essays by year one native speakers of English. BAWE-D is this sub-corpus, split into the disciplines of arts and humanities, and life, physical and social sciences.

Bi-gram
A pair of adjacent words that collocate significantly. Functional bi-grams often consist of frequently co-occurring prepositions and articles, such as *of the*, while lexical bi-grams include collocations such as *chain smoker*.

BNC
British National Corpus: a 100 million-word corpus of speech and writing.

Brown Corpus
The Brown University Standard Corpus of Present-day American English: a corpus of written English.

Chi-square test
A measure of statistical significance.

CIA
Contrastive Interlanguage Analysis (see Section 3.1)

CJA14
Corpus of Journal Articles, 2014 version. A 6 million-word collection of articles from 721 high-impact journals in 38 disciplines.

CLEC
Chinese Learner English Corpus

Colligation
The collocation of a word with a class of grammar words; e.g. ‘yet’ colligates with the present perfect in ‘Have you done it yet?’ Words can colligate with parts of speech, tenses, voice or position in a sentence, such as initial word position for connectors.

Collocation
Words that are usually found near or together with other words; e.g. ‘draw’ with ‘conclusions’, but ‘draw recommendations’ sounds unnatural due to its infrequency, and therefore does not collocate.

Concordance A list of examples of a searched-for word or phrase with the surrounding words from the original text.

Concordancer A computer program which identifies a word or phrase within a text, and outputs instances of its occurrence with the surrounding words from the original text.

Content words Words that carry information, usually nouns, adjectives, main verbs and adverbs. Often contrasted with function words.

Corpora The plural of corpus.

Corpus A collection of texts in computer-readable form, selected according to specific criteria such as genre or type of author.

Corpus balance The range of different types of language that a corpus compiler claims that the corpus covers.

Corpus-based linguistics A research method in linguistics in which researchers ‘test existing theories or frameworks against evidence in the corpus’ (Cheng, 2012, p. 6).

DDL Data-driven learning, usually involving students using a concordancer to investigate words in a corpus.

EAP English for Academic Purposes

Effect size A statistical measure that represents how big the difference between two corpora is; e.g. for the frequency of a word or phrase.

EFL English as a Foreign Language

EGAP English for General Academic Purposes: usually EAP for a class of student from more than one discipline; e.g. with Business and Engineering students together.

ELF English as a lingua franca: English when it is used by non-native speakers of different mother tongues to communicate.

ESAP English for Specific Academic Purposes: usually EAP for students from a single discipline or related disciplines with similar language needs.

ESP English for Specific Purposes

Error tagging Adding code tags to words in a corpus that indicate an error, for example in grammar.

Fisher Exact Test A statistical test of significance used for small sample sizes.

FLOB Freiburg-LOB corpus: a 1991 update of the London-OSLO-Bergen corpus.

Frequency

The number of occurrences of a linguistic feature in a corpus. Also sometimes called the raw frequency, and abbreviated as ‘Freq.’

FROWN

The Freiburg-Brown Corpus of American English, a 1991 update of the Brown corpus.

Function words

Also known as grammar words, such as pronouns, determiners and conjunctions.

Genre

A class of communication events (Swales, 1990, p. 45), including their discourse, participants and social environment (Nesi and Gardner 2012, p. 24). Definitions vary in the literature, for an analysis see Lee (2000), who gives ‘Essay – university’ as an example genre (p. 271).

Genre analysis

‘The study of situated linguistic behaviour in institutionalised in academic or professional settings’ (Bhatia, 1997, p. 629).

ICLE

The International Corpus of Learner English: a multi-national learner corpus. PLEC is a sub-corpus of this corpus.

Interlanguage

A learner’s knowledge of the language that he/she is learning. It most probably contains misconceptions about the language that he/she is learning, and gaps in the knowledge.

Keyword

A word in a corpus that occurs with a different frequency to the same word in another corpus. It shows what the corpus is about, for example corpora containing student essays on the same topic will have keywords related to that topic.

KWIC

Key word in context (unrelated to ‘keyword’ above). The key word is the word or phrase that a concordancer user is searching for, and KWIC is the concordancer output of examples of that word. The context is the other words that surround the key word, for example, the sentence that contains it.

Learner corpus

A corpus composed of texts written by language learners.

Lemma

A group of words with the same stem, such as *talk*, and belong to the same word class; e.g. *talk, talked, talking* and *talks* are all verbs.

Lexical bundle

A continuous group of words, usually three or four, found by software to collocate significantly.

Lexicon

An inventory of words in the same language. A learner lexicon is the words that a learner of a language knows.

LOCNESS

The Louvain Corpus of Native English Essays, written by British and American writers.

Log-likelihood
A statistical measure used to assess the probability of two variables being related by chance or not.

MAT
Multidimensional Analysis Tagger – a piece of software by Nini (2014) that automatically tags English texts for multidimensional functional analysis (Biber, 1988).

MI
Mutual Information: a statistical formula regarding the relationship between variables.

N-gram
A group of words that are not necessarily continuous or in the same order, usually three or four words long, and found by software to collocate significantly.

Normalised frequency
The frequency of a word or phrase divided by the number of words in the corpus, and multiplied by a factor such as 100 to give a percentage, or a higher multiple of 10 that gives an easy-to-understand and easy-to-compare number, such as an integer instead of a decimal. This allows better comparison between word frequencies in different corpora.

Parsing
Annotating a corpus for sentence structure, such as noun and verb phrases, usually done automatically by software called a 'parser'.

Pedagogic corpus
A corpus compiled from texts relevant to the teaching of students, for example from textbooks.

Phraseology
Lexical and grammatical features involving phrases related by factors including collocation, colligation, and semantic prosody.

PLEC
PolyU Learner English Corpus: a learner corpus of academic essays written by first-year undergraduates taking an EAP subject. PLEC-D is this corpus, split into the disciplines of arts and humanities, life, physical and social sciences. PLEC-EAP is the PLEC corpus, split into sub-corpora of different grades (from A+ to F) given for the essay assignment that generated the corpus texts.

Range
A technical term for the frequency of occurrence of a word that appears in a word list across a range of disciplines or sub-corpora of a corpus. Used to establish that the word is worth knowing in order to comprehend a variety of texts.

Reference corpus
A corpus that another corpus is compared to, usually a less-specialised one containing a greater range of disciplines and genres.

Semantic prosody
The collocational meaning arising from the interaction between a word and its collocates. For example, the word *cause* usually has a negative semantic prosody because it collocates with something unpleasant or bad; e.g. the cause of the crash.

SLA
Second Language Acquisition: learning a second language.

Specialised corpus A corpus that is discipline or genre specific; e.g. corpus of engineering texts or a corpus of research articles.

Sub-corpus Part of a larger corpus, usually with a feature that distinguishes it from the rest of the corpus; e.g. a sub-corpus of essays within a corpus of academic writing.

Tag Information added to a corpus text in order to label various parts or aspects. For example: the_DT shows *the* tagged as a determiner.

Tagger A computer program that adds tags to a corpus automatically, for example by deducing the part of speech for each word and adding a part of speech tag to it.

Tagging Similar to annotation: adding tags to a word in a corpus that give more information about it; e.g. 'professionals_NNS' is tagged as NNS, meaning 'plural noun'.

Tagset A set of standard tags.

TL Target Language: the language that a language learner is learning; e.g. English.

Token An occurrence of any word form. High-frequency words have multiple tokens in a corpus because they occur multiple times. The total number of tokens in a corpus is the same as the total number of words.

T-test A statistical test used to check if two sets of data are statistically different from each other.

Type A word form. Each type occurs once on a word frequency list (see the entry for word frequency list below)

Type-token ratio (TTR) A measure of lexical diversity, for example used to measure if learners know a range of forms of the same headword. Standardised TTR is measured for groups of words; e.g. every 1,000 words, therefore avoiding different text lengths influencing the TTR.

Word frequency list A list of different words, usually all the words in a corpus, together with their rank and frequencies. The word 'The' is usually ranked number one, because it has the highest frequency. Not to be confused with a word list, such as the Academic Word List.

Wordsmith Computer software for concordancing, producing word lists, managing corpus files, and related operations.

Sources: McEnery, Xiao and Tono (2006, pp. 344-351); Flowerdew, (2012, pp. 320-324); Wichmann, Fligelstone and McEnery (1997, pp. 323-326).

Chapter One: Introduction

This chapter explains the background to the research field, the research gap, the significance of the study, and the research objectives. It also delimits the scope of the research.

1.1 Background

In English-medium universities around the world academic essay writing skills are taught to students, who then are assumed to utilise these skills to write academic essays as part of their studies in the other subjects of their degree courses. However, research has shown that 'academic writing is the principal source of difficulty for Hong Kong undergraduates' (Evans and Green, 2007, p. 10). Teaching and learning materials are developed to facilitate the development of students' academic essay writing skills, and sometimes these materials are based on analyses of large collections of texts known as corpora (Hyland, 2015a, p. 203).

1.2 Research Gap

However, a key question is whether these academic essay writing skills are rewarded when students transfer them to writing for their other subjects. The teachers of these other subjects may be mainly concerned with subject knowledge, and may not assess language issues beyond the level of comprehensibility (Leedham, 2014, p. 115).
Evans and Morrison (2012) state that 'content-area professors take little or no account of English skills when assessing students' assignments, which raises doubts as to whether university English courses serve any useful purpose at all' (p. 21).

However, Hyland's (2015b) research shows the importance of the teaching of writing: 'although faculty would like to see their students write in disciplinary approved ways, their feedback rarely supports this, while students often take negative messages from the feedback concerning their learning, disciplinary communication and teacher-student relationships. As a result, EAP writing teachers are often the only resources students have in acquiring a better understanding of writing and its relation to disciplinary practices.'

Hyland (2008b) also points to a research gap when he recommends that the further study 'can offer insights into a crucial, and often overlooked, dimension of genre analysis and help provide us with a better understanding of the ways writers employ the resources of English in different academic contexts' (p. 20).

Therefore the research gap that this thesis seeks to fill is to understand how university-level English subjects can be improved to help students in their studies. It does this by examining a corpus of expert writing that gained merit or distinction grades, and was written in good English as it was written by university-level native speakers. This corpus is compared with the writing of Hong Kong undergraduate learners of English, using findings from the literature of corpus linguistics as the source of comparisons. The differences between the corpora are then used to make suggestions for the content of English subjects, with the aim of improving the English subjects and the students' writing.

This research gap can be described by the five criteria for an intellectual problem that are set out in Dunleavy (2003, p. 23), and used to define a topic in French doctoral education. Firstly, it has a goal or objective that can show that an improvement has been achieved. In this case the research will show that a corpus-based analysis of students' writing can show how lower-level writing in a corpus of writing by learners can be improved to more closely resemble the better writing in the high-grade or expert corpus, although taking into account the different contexts of the writing.

The second criterion is that there should be an initial state composed of a starting situation and resources to be used. In the case of this research, the starting state is the performance of the students as shown in the learner corpus, and the resources are the literature and tools of corpus linguistics.

Thirdly, there should be a set of operations that can be used to change the initial state. In this research, these operations are the stages of an approach called Contrastive Interlanguage Analysis (CIA) (Granger, 1996), and are detailed in Chapter Three on the conceptual framework of the study. In brief, the steps of CIA consist firstly of the collection a corpus of advanced non-native English essay writing, and secondly comparison with a control corpus of comparable writing by native speakers of English (Granger 1996, pp. 43-6). This comparison is aimed at highlighting differences between the corpora, upon which suggestions can be made for new language learning materials, which is the final stage, for reasons that are elaborated below.

The fourth criterion examines the constraints, and designates inadmissible kinds of operations. In this study the constraints include the availability of suitable well-reputed corpora, as compilation of a Hong Kong equivalent to the reference expert corpus is beyond the resources of this thesis. Therefore, the PLEC corpus is used, because it is part of, and meets the standards of, the internationally well-known ICLE corpus. The inadmissible operations are the analysis of essays by native speaker intuition, which are replaced by empirical, quantitative, corpus-based research, although the intuitions and knowledge of the native-speaker author are allowed in the interpretation of the findings (McEnery, Xiao and Tono, 2006, p. 7). This is necessary due to a second constraint, which is that the quantitative data needs to be interpreted into usable information. This is because, as Szudarski (2018) states, computers ‘cannot explain why a given feature is used in a specific way’ (p. 10). The quantitative data and its interpretations can be found in Chapters Four and Five. For example, if the learners are over-using a language construct such as personal pronouns in comparison to the expert native speakers, the reasons for this should be investigated before methods of handling the issue can be recommended, rather than, for example, taking the statistics that learners are over-using them, and immediately recommending that they be used less. This is because there may be a good reason that the personal pronouns are used more, such as instructions to give a personal opinion in the students' writing task for the learner corpus essays. As Szudarski (2018) explains, especially in the case of specialized corpora, such as the ones used in this research, 'if the analyst knows a given context in which the data are collected, they are able to account for how contextual features such as the setting, text type and communicative purpose have a bearing on the use of specific linguistic features' (p. 10).

The fifth and final criterion is that the problem should have an outcome, in that the initial state has been changed by the application of the set of operations in a way that does not violate the constraints, and meets the goal set out in the first criteria. In this research, the outcome consists of recommendations for new inclusions of input into academic writing programmes, which are made in Chapter Six. However, because the content of such programmes should be based on many factors, including learners' needs, teaching objectives and teachability (Granger 2015, p. 19), this research does not go beyond such recommendations, as programme and subject leaders are best placed to include and prioritise the teaching and learning needs and materials for their students.

1.3 Significance of the Study

The results of the study will focus on what, if any, features of academic essay writing are valued by university staff, as shown by comparing a corpus of high-scoring essays of discipline-specific topics written by native speakers of English against essays at a range of grades written by non-native speakers, and analysing what may be useful to teach. This will be able to be used by developers of EAP academic essay writing materials, and be of particular relevance to Hong Kong universities, in which thousands of students are taught these skills.

In addition, aside from the practical concerns of what can usefully be taught, there are issues of student satisfaction with educational outcomes. English language ability is important to students. In Evans and Morrison's (2012) study, graduating students are cited as being 'far from satisfied with their English skills on graduation, lamenting… their unsophisticated writing style, limited repertoire of sentence patterns and imperfect mastery of grammar' (pp. 40-1).

Therefore a major significant outcome of the study involves closing the gap between on the one hand the desires of the students and their English teachers for better academic English, and on the other hand the expectations of the content teachers and the features that they reward in their students' English.

That closing this gap is pedagogically possible is an issue addressed by Hyland (2008b, p. 4). He advocates the teaching of genre-specific bundles, in which bundles are frequently-occurring word sequences that are recurrent in that they recur at least ten times per million words and across five or more texts, and genre-specific bundles are the more frequent fixed phrases of a discipline. He raises the possibility of 'encouraging learners to notice these multi-word units through repeated exposure and through activities such as matching and item identification. Consciousness raising tasks which offer opportunities to retrieve, use and manipulate items can be productive, as can activities which require learners to produce the items in their extended writing' (p. 20).

More detailed pedagogical suggestions are given by Liu (2012, p. 33), who recommends that fixed multi-word constructions such as *in terms of* may be learned in unanalysed chunks, whereas unfixed ones in which the word order differs or different words may be inserted into the construction, such as *take / be taken into account*, should be analysed so that students can use them accurately in production.

Genre-specific and discipline-specific analysis are supported by Gardner (2012), who, commenting on her analysis of the British Academic Written English (BAWE) corpus, states that 'In Essays students develop arguments in discipline specific ways. Thus the Economics question would be answered differently if it were set in a Politics course, just as the Sociology question would be answered differently in a Law course. It is not only that they draw on different theories and construe evidence differently, but the way claims are made on the basis of evidence also differs' (p. 2). She and Nesi also found that academic staff who teach content subjects 'felt it was the subject area's responsibility to introduce students to norms specific to their area, irrespective of norms in other areas' (Nesi & Gardner 2006, p. 114), thus highlighting the existence of these specific norms, and the need to analyse corpora in these terms. A specific example of such norm-specific vocabulary is described by Breeze (2011), who found that *reasonable/ly, appropriate/ly, correct/ly* and *proper/ly* appear to convey attributes that have particular importance in the legal profession. The significance of this is the move away from the analysis of entire corpora or corpora of multi-genre academic writing, for example in Liu's (2012) multi-corpus study, towards the analysis of genre-specific corpora and sub-corpora to detect the realisations of these argumentative functions.

This thesis is also significant because it applies for the first time a number of corpus linguistics tools to the corpora, including tools in the area of phraseology. One of these is ConcGram, a tool which is used to find and analyse lexical bundles that vary in word order and multi-word bundles that vary in separation between components. A search of published papers that utilise the BAWE corpus has resulted in no example being found of the corpus having been analysed using the ConcGram software. In their introduction to ConcGram, Cheng, Greaves, Sinclair and Warren (2008) opine that ‘the over-reliance on single word frequency lists and key words needs to be redressed by examining the phraseological profile of texts, specialised corpora, and general reference corpora’, and this study will go some way towards that goal. In addition they state that the use of concgram analysis will have ‘an impact on the learning and teaching of vocabulary which will need to be addressed if phraseology is to receive the attention it deserves in language syllabi’ (p. 250), and this study will encompass both the analysis of corpora and the utilisation of that analysis. The findings of this analysis can be seen below in the Phraseological Profiles section of Chapter Five.

The originality of this study can be assessed by the application of the University of London's criteria, which are that the research should either report the discovery of new facts, or demonstrate the ‘exercise of independent critical power’ or both (Dunleavy 2003, p. 27). This study attempts to do both, firstly in that the new facts will be new differences found between the performance of the two groups of students, as shown by the comparison of their work in the corpora. Secondly, the ‘exercise of independent critical power’, interpreted by Dunleavy as the ability of the author to marshal ‘some significant theoretical or thematic arguments in an ordered and coherent way, and can explore already analysed issues from some reasonably distinctive angle or perspective’, will be attempted by exploring the already-analysed issue of the differences between Chinese and English students’ academic essays from the distinctive angle of comparing an as-yet mostly unanalysed group of Hong Kong Chinese students' essays, with a group of similar essays written by British students: a comparison that has not been done before, so providing the distinctive perspective. The theoretical and thematic arguments are laid out in Chapter Two, and their effect on the application of the findings of the research is discussed in Chapter Six.

1.4 Research Objectives

The comparison of essay writing in the two corpora, based on suggestions in the literature, aims to lead to recommendations for new input in academic writing programmes. The assumption is that current academic writing courses, which are frequently based on a needs analysis that includes examination of good models of academic writing, could benefit from corpus analysis which can extract features of these good models. These features may be generic, and therefore possible to derive by comparison of all the BAWE essays in contrast to PLEC, or they may be discipline-specific, in which case the analysis will be of sub-corpora. It has been possible to create discipline-specific sub-corpora for both BAWE and PLEC because the essays within them are organised by discipline or academic department.

The overall objective of the research is to help improve students' writing. To this end, the findings lead to a discussion of whether and how the results of the analysis can be implemented in teaching, such as in curriculum and materials design, or learning, for example through the use of concordancers in class. This forms the focus of Chapter Six: Discussion.

1.5 Scope of the Research

The scope of this research is limited in terms of the linguistic background of the student writers, by corpus and by research methodology. The linguistic background is limited because only students in two educational systems are compared: the British students and the Chinese students in Hong Kong. These limitations are based on the choice of corpora. The BAWE corpus provides a source of essays written by native-speaker tertiary students, and provides a model of what can be expected from university students writing in English. The assumption here is that it forms a model, and one piece of evidence for this comes from the readability scores of the essays in the two corpora, with the essays written by year one English speakers in the BAWE of about four grades (years of study in the American education system) above the PLEC essays (see Section 4.4.9).

Another piece of evidence for using BAWE as a model is that it is designed to contain high-quality texts. Gardner, one of the authors of the corpus, states that 'Our aim was that the Corpus should consist of good written assignments from many different disciplines and indeed from across universities. To operationalise this, we assumed that assignments which had been awarded good marks by subject tutors would qualify as well written in the disciplinary communities.' (2008, p. 3)

A British corpus was selected because the official university medium of instruction of the university at which the PLEC corpus essays were collected is British English. It is thus assumed that the university's academic staff, for example those in Hong Kong, who might be most interested in this comparison, will value similar language use to the academics who assessed the BAWE students' writing due to both the medium of instruction and similar disciplinary communities. Unfortunately there is no existing corpus of native-speaker, high-graded essays written by Hong Kong university students, that would show both the language and content expertise which would be necessary as proposed models of performance. Such a corpus would be challenging to collect sufficient essays for, due to the lack of native speakers of English in Hong Kong universities, thus restricting the corpus size to a level that might bring into question its discipline coverage and representativeness.

It should be emphasized that a corpus of research articles from academic journals was deliberately not selected as the main comparison, as has been done in other research, for example Bloch's (2010) corpus of critical book reviews and academic reports from the journal *Science*, because research articles are both a different genre of writing and are written by authors with more training and experience in academic writing skills, as well as being edited. That the genre is different can be seen, for example, in the organization of research papers, which usually contain sections such as the abstract and the literature review that are often not present in academic essays. That the authors generally have more training and experience of academic writing can be seen from their qualifications, usually at a master's or doctoral level, and the experience of writing the theses necessary to attain these qualifications. Hyland and Milton (1997) comment that students 'are often measured against an unrealistic standard of "expert writer" models such as academic research articles, a genre which is typically rigorously reviewed and revised before publication' (p. 184).

However, a corpus of academic articles is used in this thesis in contexts in which comparison to such a corpus is appropriate. This is the Corpus of Journal Articles 2014 (CJA14) (Research Centre for Professional Communication in English, 2014) which is a 6,015,063-word collection of articles from 721 high-impact journals in 38 disciplines. There are a number of reasons for using the corpus, including that, as Lee and Chen (2009) point out, having 'two corpora to compare our learner data against gives us added confidence that we are focusing on the right words' (p. 154).

The research methodology also affects the scope. This thesis is corpus-based in that it uses corpus linguistics ‘to test existing theories or frameworks against evidence in the corpus’ (Cheng, 2012, p. 6). It utilises a range of software to analyse the corpora on a larger scale than that which would be possible by manual analysis, although this does limit the interpretation to that of what the software reveals, which may be less that what a manual analysis could show.

There are no interviews with teachers or students. This is because, for the teachers of the students in the PLEC corpus, they may feel pressure from the university English language policy that the medium of instruction must be in English unless given a special exemption (for example for Chinese language subjects). Their students are unlikely to know what aspects of their English the teachers value and reward, if any, in their work. In addition, neither the BAWE or PLEC are live corpora, and tracking down and interviewing the teachers and students involved would be impractical. However, both the works of Leedham (2014) and Durkin (2011) do contain the findings of interviews with students, and they are included where appropriate.

Due to the breadth of the literature on this topic, there is a need for principled selection of research to include for analysis in the Findings chapter. The principles are that either the research should be widely-cited in the literature, such as the Academic Word List (Coxhead, 2000), it should be related to BAWE or PLEC, or related to the needs of Hong Kong tertiary students and their academic essay writing.

1.6 Summary

This chapter has explained the background to the research, that academic essay writing is a common task for tertiary students, and that corpus linguistics can be of assistance in writing materials for these students. It has also identified the research gap, that the design of such materials can be informed by corpus-based analysis of both the language that such students produce and comparison with a similar analysis of the language produced by native speakers of an equivalent stage in their university careers who are writing a broadly similar text-type. This research gap was found to fulfil the criteria of a doctoral education intellectual problem from Dunleavy (2003).

The significance of the study was then addressed in terms of the large number of students affected and the language issues that these students face. The originality of the research was then assessed and found to fulfil the criteria of discovery of new facts and exercise of independent critical power. Following on from this, the research objectives were explained, in that a corpus-based analysis could suggest areas of student writing which could be improved, thus adding new knowledge to the field. Finally, the scope of the research was delimited to a comparison of Hong Kong and British university students' academic essays, with supporting information from other corpora where appropriate.

The next chapter follows on from this by providing a review of the literature of English for Academic Purposes and corpus linguistics, narrowing to a review of similar research that compares Chinese and British students essay writing, albeit at different groups of Chinese and British students of slightly different ability levels. It also details the corpora used in the study, and previous research based on them, as well as suggesting how this research may contribute to the literature through its analysis of two hitherto rarely-compared corpora.

Chapter Two: Literature Review

This chapter details previous research in the field and how this leads to the current research, describes the corpora that are used in this study, and explains how this study can contribute to corpus linguistics.

2.1 Previous Research

This section examines previous research on English for Academic Purposes, followed by research that compares British and Chinese students' writing.

2.1.1 English for Academic Purposes

English for Academic Purposes (EAP) is defined by Jordan (1997) as 'concerned with those communication skills in English which are required for study purposes in formal education systems' (p. 1). He divides the purposes of learning English into general, social and specific (ESP), and sub-divides ESP into English for occupational, professional, vocational and academic purposes (EAP). He goes on to further sub-divide EAP into English for Specific Academic Purposes (ESAP), for example for medicine or engineering, and English for General Academic Purposes (EGAP), that includes academic writing, in which he includes skills such as academic style and proficiency in language use (1997, p. 3). Regarding language use, he states that overseas students that he studied said they had problems in, from most to least common, vocabulary, style, spelling, grammar, and punctuation, and that academic staff commented that they had most difficulty with (in order) their students' style, grammar, vocabulary, punctuation and spelling (1997, pp. 46-7).

Overseas students are often non-native speakers of English. They may or may not have academic study skills from their own language (Jordan 1997, p. 5), and for those who do, the academic conventions from their native environment may be different from those expected in the English-speaking environment (Durkin, 2011). Hyland (2015a, pp. 48-60) points out that English is increasingly being used in international academia, and that teachers ‘frequently reject non-standard varieties of English’. He further points out that in EAP students are expected to develop academic literacy, which is not a single literacy, nor control of grammar, nor the ability to transform knowledge, but the ability to write as a member of a discourse community (2015a, p. 39).

An example norm of writing in the academic discourse community is the difficulty of reading the text that has been written (Hartley 2008, p. 5). Computer-based measures of text difficulty can be carried out by examination of sentence structure, in terms of word and sentence length, which can be measured by readability scores such as the Flesch Reading Ease score. These scores predict the number of years of education of the writer, with student essays and academic articles being the most difficult (Hartley, 2008, p. 7).

That there are differences between the academic essay writing of British and Chinese students, of whom Hong Kong students form a sub-set, was demonstrated by Leedham (2011). Based on a comparison of Chinese and English native speakers in the BAWE corpus, she found that, for example, the Chinese students make greater use of particular connectors and the first person plural (p. 1). This observation seems to be generalisable, as Adel and Erman (2012) conclude that there is a general pattern found in research that ‘non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers’ (p. 90), even when the non-native speakers are of an advanced level. They found that native speakers tended to use more unattended ‘this’ constructions, existential ‘there’ constructions, hedges, and passive constructions.

However, these differences may depend on the language features of the non-native speakers’ first language. For example, Lin (2003, p. 284) describes how Chinese has a similar construction to existential ‘there’ constructions. This influences the learner’s interlanguage. Larsen-Freeman and Long (1991, p. 60) describe interlanguage as a continuum between the L1 (first language) and L2 (second language), that all learners traverse in a systematic way, and Huat (2012) states that this learner language ‘constitutes a linguistic system in its own right rather than being a deficit version of that of an idealized monolingual native speaker’ (p. 195). Thus Adel and Erman’s finding needs to be checked in regard to the Hong Kong students’ language production, because the general pattern may not be applicable in their case.

2.1.2 Comparisons of British and Chinese Students’ Writing

Differences between the academic essay writing of British and Chinese students have been investigated before in the research literature, especially using the BAWE corpus, but comparing it to different corpora from those in this research.

Chinese students’ writing was investigated in a corpus-driven study by Leedham (2014). The two corpora that she used were both sub-sets of the BAWE corpus: the reference corpus consisted of 611 assignments, comprising over 1.3 million words written by 70 students for whom English was their L1, and the learner corpus contained 245 Chinese students’ assignments. She also used extra assignments, which were collected and examined in the same way as BAWE. Her final corpus used was termed ‘Chi123’ because it contained texts from first, second and third year Chinese students. It consisted of over 279,000 words by 45 students. These figures contrast with the corpora used in this study, which are both composed by first-year writers’ only, the text-type is essays only, and PLEC is larger than Chi123, making my corpora of broadly similar word count. It should be noted that in the BAWE corpus, there are 66 texts labelled as Chinese Cantonese, 26 as Chinese Mandarin, and 153 as Chinese unspecified. This, and the fact that in the category ‘Education’, non-UK students are all labelled as ‘OSA’ or ‘Overseas All’ in the BAWE Manual, make it difficult to know how many students are from the Hong Kong educational system, and how many from the separate mainland Chinese system.

Leedham's approach was corpus-driven, using keyword analysis to reveal unexpected patterns, rather than using the corpus-based approach in this thesis, which takes existing findings from the literature and examines the corpora in their light. Her research questions involved the distinguishing characteristics of the writing in the Chi123 corpus, variation between writing by students in different years, and the effect of discipline on writing.

Another study which used the BAWE corpus was by Chen and Baker (2010), who used it for their investigation of the relationship between learner proficiency and n-grams, which are statistically significant contiguous stretches of words that are discovered computationally. The difference between their research and this thesis is that the Chinese and English learner corpora that they used were both from BAWE and were not confined to essays, and their expert corpus was an academic prose sub-set of the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus. Their results and similar results from the PLEC-EAP, BAWE-EON and CJA14 corpora are analysed and discussed in the n-grams part of the Vocabulary sections in the Findings and the Discussion in Chapters Four and Six below.

Chen and Baker's paper, along with over one hundred other corpus research papers which have used the BAWE corpus, are listed on the BAWE website (Research using the BAWE corpus, 2018). Those relevant to this research are detailed in the Findings chapters below.

One other work especially relevant to this thesis is Milton’s (2001) research report on the elements of a written interlanguage, which is a corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students. This report investigated a corpus of Hong Kong learners Use of English ‘A’ level school-leaving exams containing argumentative and discursive essays from the early 1990s, and compared them to high-scoring British students ‘A’ level exams of the same genre, thus providing a snapshot of students’ writing a few months before some of them, probably about the best fifth of them in terms of exam scores, came to university. Milton examined the students’ interlanguage in detail, and relevant findings from his work are referenced in some of the chapters below.

To understand the context of the research detailed above, it is necessary to know more about the corpora involved, and that is the topic of the next section.

2.2 Description of the Corpora

This section describes the two main corpora that are contrasted in this study: an ‘expert’ corpus and a ‘learner’ corpus, and a number of other corpora that are used as supporting evidence or for contrast.

2.2.1 The BAWE Corpus

The BAWE corpus is used as an ‘expert’ corpus, which is defined as a corpus that serves as a benchmark ‘against which learner production can be assessed’ (Szudarski, 2018, p. 117). Texts for the corpus had been positively assessed by subject tutors, and it contained only merit and distinction level work (Nesi and Gardner, 2012, p. 7). It provides both a model of good student-level language use and content that meets the approval of academic staff (Flowerdew, 2012, p. 171). Leedham (2014) comments that ‘access to proficient student writing corpora, such as BAWE, … are of great potential value … in that they can be seen as representing target writing’ (p. 129).

The corpus consists of 2,897 student assignments from three British universities: Oxford Brookes, Reading and Warwick, collected in a project that lasted from 2004-7. The assignments are from 35 disciplines, from Archaeology to Sociology, covering arts and humanities, life, physical and social sciences (Heuboeck, Holmes & Nesi, 2010). The total number of words is just over 6.5 million, fairly evenly distributed between the disciplinary groups. The texts are from students in years one to three, and master’s level, again fairly evenly distributed. They range in length from about 500 to 5,000 words. The assignments were divided into 13 genre families, and the majority came from the essay genre family, with 1,238 texts. They were not evenly distributed in terms of discipline, with 602 coming from arts and humanities, 127 from life sciences, 65 from physical sciences and 444 from social sciences. Students assigned copyright of their assignments to the universities concerned, and provided meta-data which was used in this research to filter out assignments, such as those from Chinese speakers.

In order to provide a more appropriate comparison to the PLEC corpus, in which the essays are only written by first-year undergraduates, a sub-corpus of BAWE called BAWE-EON was created for this study which isolated the academic essays from year one native speakers, with BAWE-EON standing for BAWE Essays by year One Native speakers. This sub-corpus contains 330 essays comprising 640,013 words, 24,572 s-units (sentences counted by a sentence-splitter algorithm, rather than manually), and 4,570 p-units (paragraphs). The discipline composition of these 330 essays varies, with the top 3 most frequent disciplines being humanities subjects: Philosophy with 35, and both Classics and English with 30. Sciences were less frequently represented, with Agriculture, Biological Sciences, Chemistry and Engineering all represented by only one essay. The BAWE documentation divides the essays into disciplinary area, and in the sub-corpus, 198 essays were in the Arts and Humanities area, 34 from Life Sciences and Medicine, 21 from Physical Sciences and 77 from Social Sciences. This disciplinary-categorised sub-corpus of BAWE-EON will be referred to as BAWE-D, and is used in the research on disciplinary variation below.

The BAWE corpus was chosen because it is suitable for contrast to PLEC in that it is an expert corpus, which has been used in other research (e.g. Leedham, 2014; Li, 2014; Lee and Chen, 2009) as an expert corpus, was written by students at a similar stage of their university careers, and in the same language variety, British English, as the PLEC corpus. BAWE is designed to be used for comparison with other corpora, specifically the Michigan Corpus of Spoken Academic English (MICASE) and the British Academic Spoken English (BASE) corpora, and it is thus very well documented in the BAWE manual (Heuboeck, Holmes & Nesi, 2010), which details the files contained and allows the creation of sub-corpora. BAWE-EON is of a similar word count to PLEC.

2.2.2 The PLEC Corpus

The PLEC corpus is a learner corpus which contains over 1,200 student argumentative essay assignments from 22 university departments, from accounting to logistics, totalling just under 670,000 words. The assignments are by non-native speakers, and quotes have been removed to ensure that the texts are all the students' own work. All the assignments are of the university argumentative essay genre type. PLEC was accepted as a part of the International Corpus of Learner English (ICLE) (Granger, 1990), and as a part of ICLE, it follows the ICLE guidelines that the texts should be argumentative essays of 500 – 1,000 words, on non-disciplinary topics, with students' spelling errors left uncorrected, and with references and quotes removed. The essays were written in one hour 45 minutes under exam conditions, based on a given topic of recycling, banning smoking, credit cards, Hong Kong country parks, immigration from China, Lowu railway, peer assessment or cyber cafes, and six short paragraphs of background material on each topic. To reduce the cognitive load caused by the time limit, students had discussed the topics before the timed writing.

These features make the use of PLEC from a corpus linguistic point of view, and for comparison to BAWE, a little more problematic, as the spelling errors can interfere with concordance searches, the BAWE essays were not timed, and the BAWE essays were in the students' disciplinary area. The effect of this, as Leedham (2014) points out, is that 'It could be the case that ICLE contributors would write differently in academic essays on an academic topic area they were familiar with' (p. 33). However, PLEC students, mostly being about two months into their university careers, and many of whom were studying specialised academic fields such as engineering that they had not studied at secondary schools, probably had a limited level of expertise in their discipline.

PLEC is divided into three sub-corpora, which organise the texts in various ways. The PLEC-EAP corpus organises the texts by the grades that the students achieved on an English for Academic Purposes subject, and this corpus that is used in this thesis when comparisons of students' ability are required. It contains 657,339 tokens according to WordSmith 7, in 1,271 essays. There is also a PLEC UE corpus, in which the essays are organised according to the student's grade on the Hong Kong public examination called Use of English 'A' level. Students would have normally taken this exam at the end of their secondary school careers, usually about 5 or more months before they wrote the essays that are in the PLEC corpus. In those months the student took the EAP university subject, so the PLEC EAP grade is a more up-to-date assessment of their English abilities, and is a test of the academic writing skills that they had not yet been taught when they were taking the UE exam.

The third categorisation of the PLEC students work is by their university department. A discipline-specific version of the PLEC corpus called PLEC-D was created by categorising the files from students of various departments into similar categories to those used in BAWE-D, and these are shown in Table 2.2 below.

In this research, the PLEC corpus can be regarded as both a learner corpus, in that it is a corpus written by language learners, and a pedagogically-oriented corpus, which is defined by Szudarski (2018) as a 'small-scale learner corpora that consist of the language produced by L2 learners', which 'can be used for identifying the most frequent errors learners tend to struggle with. This in turn can lead to corpus-informed form-focused instruction that targets the most problematic aspects of L2' (p. 109), which is the objective of this thesis.

The BAWE-EON sub-corpus and the PLEC corpus were annotated with part-of-speech tagging using TagAnt 1.2.0 (Anthony, 2015). Meta information in the corpora, such as <a> indicating a grade A essay in the PLEC corpus, which the tagger had tagged as an article, was then untagged to avoid interfering with the counting of parts of speech instances. Similarly in the BAWE-EON sub-corpus, meta information such as headings, which the tagger had converted to <_SYM heading_NN >_SYM, was reverted to <heading> to avoid them being counted as a noun.

Two other corpora are used in this study to provide additional information. This follows the model of Milton's (2001) corpus-based study of the acquisition of a written English interlanguage by Hong Kong Chinese students, in which the additional corpora were used to check the observations found in the main corpora, and also follows the model of Chen and Baker (2010), who used two learner corpora and an expert corpus of published academic text.

2.2.3 The Corpus of Journal Articles 2014

In order to provide a basis of comparison with academic writing in journals, the Corpus of Journal Articles 2014 (CJA14) (Research Centre for Professional Communication in English 2014) is used. It is a 6,015,063-word collection of articles from 721 high-impact journals in 38 disciplines in Journal Citation Reports (JCR) or in SCImago Journal Rank (SJR). For this research, a sub-corpus split along disciplinary lines, and called CJA14-D, is also used, as can be seen in Table 2.2 below. The use of this corpus in some part addresses Widdowson's (2000) concerns as to the limitation of corpus linguistics, that of a written corpus only having the

Table 2.2: Disciplines in the PLEC, BAWE and CJA14 disciplinary corpora

Discipline	BAWE-D Fields	PLEC-D Departments	CJA14-D Journal Articles
Arts and Humanities	Archaeology Classics Comparative American Studies English History Linguistics Philosophy	Chinese and Bilingual Studies School of Design	Archaeology Linguistics Design Education History History of Art Literature Music Philosophy Communication
Life Sciences	Agriculture Biological Sciences Food Sciences Health Medicine Psychology	Applied Biology & Chemical Technology Nursing and Health Sciences Optometry and Radiography Rehabilitation Engineering Centre Rehabilitation Sciences	Applied Biology & Chemical Technology Health Technology and Informatics Nursing Optometry Psychology Rehabilitation Sciences
Physical Sciences	Architecture Chemistry Computer Science Cybernetics & Electronics Engineering Mathematics Meteorology Physics Planning	Applied Maths Applied Physics Building and Real Estate Building Services Engineering Civil and Structural Engineering Electrical Engineering Electronic and Information Engineering Institute of Textiles and Clothing Manufacturing	Applied Mathematics Applied Physics Building and Real Estates Building Services Engineering Civil and Structural Engineering Computing Electronic and Information Engineering Geography Industrial and Systems Engineering Land Surveying and Geoinformatics Mechanical Engineering Textiles and Clothing
Social Sciences	Anthropology Business Economics Hospitality, Leisure and Tourism Management Law Politics Publishing Sociology	Accounting Applied Social Sciences Business Hotel and Tourism Management Shipping and Transport Logistics	Anthropology Accounting and Finance Applied Social Sciences Economics Hotel and Tourism Law Logistics Management and Marketing Politics Sociology

ability to show what people write, not what they could write, or should write. CJA14 is a large, expert corpus, and has the potential to show what the BAWE and PLEC students could have, or should have, written. This is supported by Flowerdew (2012, p. 171), who comments on BAWE students' lack of genre mastery.

The reason for using the main corpora, PLEC and BAWE-EON, which are collections of the specific genre of academic essay, rather than academic writing in general, is due to the differing purposes of the writing. Ebeling (2011) contrasts the top 50 trigrams (groups of three consecutive words) from BAWE and general academic prose. For his suggestions for further study he proposes that 'studies could also include student vs. professional writing or native vs. non-native student writing in order to investigate how salient n-grams really are and to what extent the same patterns and functions are used … by learners and native speakers.' Research has also already been carried out and a word list of high frequency word forms found in the BAWE essays was available online at Coventry University (Lexical Items, n.d.). In addition, Chen and Baker (2010) compared n-grams between expert, native-speaker learner academic writing and Chinese learner academic writing using BAWE for the learner corpora.

Examples of n-grams distinctive to the BAWE corpus essays include *allows the reader to, as can be seen, at the beginning of, by the use of, can be seen in, could be argued that, it could be argued, the beginning of the, the importance of the, through the use of, to the fact that,* and *way in which the* (Ebeling, 2011, p. 60). These n-grams comprise lexical items with grammatical, cohesive and stylistic aspects such as use of the passive voice, modal verbs, connectives and hedging; aspects which Evans and Morrison (2012) found are 'generally not included in assessment criteria and thus received less attention in the planning and production of assignments than information and ideas' (p. 41).

2.2.4 The LOCNESS Corpus

Some of the research cited in this thesis, for example Gilquin and Paquot (2008), Lin (2003) and Lu and Ai (2015), uses the Louvain Corpus of Native English Essays (LOCNESS) (Learner Corpus Association, 2014), which is another reference corpus. It is comprised of 60,209 words of British pupils' A level essays: 95,695 words of British university students essays and 168,400 words of American university students' essays. Although this corpus is used in this study, it is not the main one due to its smaller size than BAWE, and inclusion of American essays in contrast to BAWE and PLEC, which both aim at the British variety of English.

2.2.5 Corpus Analysis

A number of studies use corpora to compare native- and non-native speaker language use in the area of phraseology (Adel & Erman, 2012; Leedham, 2011; Salazar, 2008). They agree with Chen and Baker (2010, p. 44) that 'frequency-driven formulaic expressions found in native expert writing can be of great help to learner writers to achieve a more native-like style of academic writing'.

There is also agreement in the literature on the need for genre-specific analysis. Cheng (2007), in her description of the uses of the ConcGram software, states that at PolyU 'specialised texts collected from such major academic disciplines as engineering, land surveying, business, financial studies, design, tourism and hospitality management, health sciences, and so on, could be concgrammed in order to determine the discipline-specific phraseological profiles', and that this is important because 'Those who fail to communicate using the conventional keywords and phraseology of business English might be misunderstood and, as readers, might misunderstand the subtle shifts in meanings that result in particular choices' (p. 31).

Such approaches to genre are partly grammar based, in accordance with Nunan's (2007, p. 71) description of grammar as being concerned with how words are combined. He contrasts the mentalist model of transformational-generative grammar, which sees language as a psychological phenomenon, and the functionalist model of systemic-functional linguistics, which is concerned with the social dimension of language, but concludes that they both may be valid. The genre-based approach tends towards the functionalist model in that it is concerned with how language is used among the social groups of practitioners whose texts are collected to form the reference corpus. Nunan distinguishes between prescriptive and descriptive grammars (2007, p. 75), and corpus linguistics can be seen as being descriptive, in that it 'focuses on describing the way people actually use the language' (p. 76).

Another approach to analysing the corpora is based on vocabulary. The range of vocabulary of the students can be compared to vocabulary lists, for example Coxhead's Academic Word List (Coxhead, 2000a). However, there is some doubt as to whether academic vocabulary is homogenous enough to be used as a measure. Hyland and Tse (2015) are concerned that the label *academic vocabulary* 'conceals a wealth of discursive variability which can misrepresent academic literacy as a uniform practice and mislead learners into believing that there is a single collection of words which they can learn and transfer across fields' (p. 386).

However, Hyland and Tse's claim has been disputed, for example by Simpson-Vlach and Ellis (2010, p. 510), who, while recognising that disciplinary variation is important, were able to derive a common core of academic formulas that they claim transcend disciplinary boundaries. This core they derived by using different bundle lengths and cut-off frequency thresholds, which enabled them to discover a number of core bundles common to all academic disciplines.

From a practical point of view for teaching and learning, there is a need for non-discipline-specific language for English for General Academic Purposes subjects. Teachers may face a class containing students from a variety of disciplines, and thus need to teach a general EAP curriculum, rather than a discipline-specific one. In addition, students may take subjects from a variety of disciplines, because, for example, at the university where PLEC was collected there is now a general education program in which students need to take subjects from outside their department across a range of broad disciplines.

Since the majority of EAP course books and subjects analysed in this research do not break vocabulary down by students' majors, there is a risk that students may not learn the vocabulary most specific to their field. However, due to this reality of teaching EAP to classes of students from a variety of disciplines, cross-disciplinary academic language use warrants investigation.

One previous comparison of the BAWE and PLEC corpora has been found in the literature: Li's (2014) comparison of the use of first-person pronouns. Using 1,213 social science essays comprising 3.3 million words from BAWE for compatibility to her other corpora, she found that 'Chinese learners of English tend to overuse *my* and *our*, about three times more than that of native English speakers' (p. 302), and concluded that this is due to mother tongue influence and signals a sense of belonging. This thesis expands on this by comparing pronoun use in all academic essays, not just the social science ones, in Chapter Four.

Thus it can be seen from this section that the selected corpora are of suitable standard, are comparable, and that there is a precedent for their utilisation in research.

2.3 Research Contribution

Li’s (2014) paper was the only research found comparing BAWE and PLEC, but it is possible that further comparison could produce valuable insights into the differences in the academic essays in these corpora, and therefore what needs to be taught to Hong Kong students.

Other research outlined above is re-examined in this thesis to assess whether the findings are accurate in the Hong Kong context, and if so, how they can contribute to the teaching of academic essay writing in Hong Kong.

2.4 Summary

This chapter has reviewed previous research on EAP and more specifically, academic essay writing and the challenges that first-year university students face with it in their development of an interlanguage from the English that they were taught in secondary school to the requirements of academia. It is these challenges that this research aims to address. The chapter has described the PLEC and BAWE-EON corpora, as well as the other corpora used, and the reasons for their selection, in order to show that appropriate data is being analysed.

The chapter has also suggested the manner in which the current study may contribute to the literature of corpus linguistics by a corpus-based comparison of PLEC and BAWE-EON in the Hong Kong context to assess whether the findings in the literature on students' academic essay writing are applicable to Hong Kong students' academic essay writing in English.

The next chapter explains the theoretical framework used and the resultant research questions. It then goes on to detail the methods used to address these questions.

Chapter Three: The Conceptual Framework

This chapter follows on from the review of the literature in the previous chapter by explaining the theoretical framework on which the research is based, how this relates to corpus linguistics, and how this leads to the research questions. The research methods used to investigate these questions is then described, including the research strategy and design.

3.1 Contrastive Interlanguage Analysis

This thesis utilises the theoretical framework known as Contrastive Interlanguage Analysis (CIA), which establishes comparisons between 'native and learner varieties of one and the same language' (Granger, 1996, p. 43), that language being known as the 'target language' (TL). According to Li (2014), 'The CIA analysis also covers the degree of the TL behaviour of the learners influenced by their native language (NL), areas for learners to achieve native-like or non-native-like linguistic performance, and predictability of learner difficulties' (p. 304).

3.1.1 Steps of CIA

The steps of CIA are described by Granger (1996, pp. 43-6) as firstly the collection a corpus of advanced non-native English essay writing, followed by a comparison between native and non-native varieties of the same language with a control corpus of comparable writing by native speakers of English. This is done by using techniques such as word frequency comparisons, concordancing, and investigation of collocations.[1]

[1] Concordancing is the use of software called a concordancer that searches the text in a corpus and outputs a screen of example stretches of text that include a match to the search term. Collocation is groups of words, usually pairs, that 'tend to be selected with each other' (Cheng, 2012, p. 4), and are therefore found either adjacent to or within a few words of each other.

CIA is related to the aims of this thesis in that, according to Granger (2003), ‘Evidence of learner under-, over-, and misuse can help materials designers and teachers select and rank ELT material at a particular proficiency level’ (p. 543).

3.1.2 Criticisms

CIA has been criticised on two main issues, the ‘comparative fallacy’ and the English native speaker as a norm and target (Granger 2015, p. 13). The ‘comparative fallacy’ points out that if an idealised native speaker is seen as the target level of proficiency, students’ interlanguages are implicitly deficient by contrast, rather than a stage in the positive learning process of a developing interlanguage. It may also not be the case that student target language is native-speaker imitation. Granger rebuts these points, saying that ‘all the studies that compare learners of different proficiency levels are in fact based on an underlying L1 norm, as proficiency is usually assessed with an L1 target in mind’ (2015, p. 14). She also points out CIA does not just have a theoretical aim, it also has an instructional one in that it aims to inform pedagogical applications. For this study, these applications are discussed in Chapter Six.

Criticism of the concept of the native speaker as the norm of language use has centred around research into World Englishes and English as a Lingua Franca (ELF) (Huat, 2012, pp. 195-6). Regarding World Englishes, there is a large range of native-speaker varieties such as British and American Englishes, and non-native varieties as well. Granger counters this argument by emphasising that the reference corpus in CIA studies does not have to be a native-speaker corpus, and she cites the wide range of sub-corpora available in ICLE, such as Indian and Singaporean Englishes. In this study, a native-speaker corpus forms the reference corpus as the official medium of instruction of the university of the learner corpus writers is British English.

The concept of the native speaker norm has also been criticised in ELF contexts, in which successful communication is more important than imitation of native speakers (Louhiala-Salminen and Kankaaranta, 2011). A *lingua franca* is defined by Gerritsen and Nickerson (2008) as ‘a third language … that both parties are able to speak and understand well enough to communicate’ (p. 180). Granger counters this concept of a native speaker norm by arguing that not enough is known about proficiency levels in ELF to be able to identify acceptable features and lexical structures, and thus form a target language. She also points out that the term ‘reference’ when referring to the reference corpus ‘makes it clear that the corpus does not necessarily need to represent a norm’ (2015, p. 17).

3.1.3 CIA^2

Despite her rebuttals of these criticisms, Granger admits that in the decades since she proposed CIA in 1996 the field has changed, and she therefore proposes a new version of CIA, which she calls CIA^2. In this, there are terminological changes, such as to ‘varieties’ that can be compared: reference language varieties and interlanguage varieties of English. Thus a reference language variety could be the language contained in the Corpus of Journal Articles 2014, which contains papers written by non-native speakers (judging by the authors’ names, as the native language of the authors is not encoded in the corpus).

Related to the concept of native-speaker norm are the terms ‘under-use’ and ‘over-use’. However, Lee and Chen (2009) point out that ‘the term *overused* is a purely statistical term — there are no prior assumptions that all such overused items necessarily represent bad writing practices. They merely point to differences which merit further, qualitative investigation' (p. 153). Such an investigation may not give rise to a recommendation for a change, as Granger (2015) states that 'using something too much may be a perfectly understandable, even desirable, feature of learner language' (p. 19).

A similar perspective on contrastive studies is taken by Leedham (2014), who views all student writing, regardless of first language, as a process of learning how to make meanings within the academy and within the discipline. The idea of a native-speaker norm can lead to a 'deficit' perspective and a need for 'remedial' materials. She prefers an alternative academic literacies perspective, which views academic writing as a social practice involving genre, context and culture, and views 'all university students as learner writers within the academy' (p. 6).

There are a number of features of learners' interlanguage (Flowerdew, 2012, pp. 169-70), and many of these are evident in the findings listed in Chapters Four and Five below. The first is L1 transfer, in which the learner uses a form or grammatical pattern found in their L1 in the L2. This can result in over-use, such as 'existential *there*' for Chinese learners, or mis-use, such as for *On the other hand* when erroneously used to mean *in addition*, rather than correctly denoting contrast. The second interlanguage feature is general learner strategies that reduce the complexity of a language learning task, such as an assessment. These strategies can include circumlocution and avoidance, for example Hu and Gu (2015) comment on cases of avoidance of the perfect aspect by Chinese learners. The third feature is paths of interlanguage development, in which the learner improves their ability over time or with proficiency, for instance PLEC students improve their use of passive voice as their proficiency increases. The next feature is intralingual overgeneralisation, in which a learner overgeneralises, for example a grammatical rule. An example of this is the PLEC corpus is the use of the plural form *researches* after structures that colligate with a plural, such as *a number of researches, some researches*, and *many researches*. The fifth interlanguage feature is input bias, in which the input that the learner has received is reflected in their output. Examples of this include the tendency to re-use and over-use language from an essay prompt in the text (Milton, 2001) and to use connectors interchangeably without adaptation for suitability with the surrounding text, stemming from inaccurate translations (Lee and Chen, 2009, p. 288) and lack of register awareness (Conrad, 2000). The final interlanguage feature is genre/register influences, such as the inclusion of informal patterns characteristic of spoken discourse in academic writing, such as the use of rhetorical questions.

In summary, CIA^2 provides a mature and detailed framework on which this thesis can be based. One of its major tools for comparison of language use is corpus linguistics, which forms the focus of the following section.

3.2 Corpus Linguistics

Corpus linguistics is a powerful research method in CIA, due to its ability to analyse large collection of texts for specific features, and generate statistics that provide strong evidence of language use. Thus this thesis is corpus-based as it tests ‘existing theories or frameworks against evidence in the corpus’ (Cheng, 2012, p. 6).[2] The external criteria used for corpus selection in this thesis are that the mode of the text should be writing, the type of text should be student essays, the domain should be academic and university, the language varieties of the writers should be either English native speakers or learners, and the location of the texts should be English of the UK or Hong Kong. Thus they are specialised corpora (Flowerdew, 2004, p. 21; Cheng 2012, p. 34) because they have a specific purpose for compilation, they are contextualised in that they have a particular setting, participants and communicative purpose, they have a specific genre as they are composed entirely of essays, and they have specific varieties of English.

3.2.1 Design Principles

Sinclair (2004a) puts forward a number of design principles for the construction of a corpus, and these were followed in the construction of PLEC and BAWE. His first principle is that ‘The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.’ The selection of the BAWE and PLEC corpora for this study is based on the communicative function of academic essay writing in the community of universities that use British English as their medium of instruction.

[2] A corpus is defined by Sinclair (2004a) as ‘a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.’ Corpora are therefore machine-readable, authentic texts that are usually samples of language chosen for their representativeness (McEnery, Xiao and Tono, 2006, p. 5).

The use of the LOCNESS corpus would complicate the analysis due to its inclusion of American essays, which would have belonged to a different community.

Sinclair's second principle is that 'Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.' For this reason the main corpora chosen were of student work and not, for example, corpora containing academic papers published in journals, even though it might be argued that those would have provided a model of the academic writing to which students should aspire. The BAWE corpus was designed with representativeness in mind (Nesi and Gardner, 2012, pp. 6-9), and therefore contains a range of genres and disciplines.

Methods of sampling language for inclusion in a corpus are Sinclair's (2004a) next concern. First he examines 'orientation', and warns that selections from a corpus may not be as representative of language use as a whole. In this study a selection of the texts that comprise BAWE is used, but the selection is necessary to parallel the text types in PLEC, and it is not claimed that the corpora constitute a reflection of all academic essay use, or all that can be learned from a comparison of native and non-native speakers' academic essay writing.

A further guideline for the construction of a corpus is that 'Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination' (Sinclair, 2004a). The BAWE Manual clearly describes the criteria used in its construction, and the texts are tagged so that they can be filtered in order to isolate the variety of text needed. The PLEC texts are all of one language and one language variety. They are also sorted in a number of ways: by the university department that the student is from, the language ability of the student as either reflected in public examination score, or by the grade for the assignment that the texts were in answer to.

In order to avoid errors in concordances, word lists and word frequency counts, it is important to avoid including computer mark-up in the text of the corpus. Sinclair (2004a) recommends that 'Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.' The BAWE Manual describes how this is done, and three versions of each text are supplied, a plain text version in .txt format, and two versions with mark-up included. In PLEC each text is tagged at the start, for example, '<deg> <lac01 s14> <a+> <uea>' referring to a degree level student writing on assignment version 01, student number 14, with an A+ grade on the assignment, and an A grade in the public examination. Because these tags are in angled brackets and do not constitute whole words they can easily be eliminated from searches, word counts and frequency measures.

Sinclair's next guideline is that 'Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.' This is achieved in both the BAWE-EON and the PLEC corpora, as they contain whole essays. However, the PLEC essays have had the quotations removed, as they are not the students' own work, and there are two versions of the corpus, one with references included, and one with them removed. The version with references included was utilised, both because the BAWE essays include references, and because referencing was found to be a frequently-taught topic on EAP courses and in EAP books. Sinclair is correct that the samples vary in size, as the BAWE-EON texts range in length from about 504 to 5,400 words. The PLEC essays are far more homogenous in length due to the fact that they were written as timed assignments.

The representativeness of the corpus is Sinclair's next concern, and he recommends that 'The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.' This is done in the BAWE manual. Due to the fact that the PLEC corpus contains a much more limited set of texts in terms, for example, of genre, information about it is more limited, although the compiler kindly sent the author a background document.

In addition to representativeness, Sinclair also advocates 'balance'; i.e. that 'the proportions of different kinds of text it contains should correspond with informed and intuitive judgements.' The BAWE manual contains tables of text types and word counts per discipline group and genre family. For PLEC there is only one genre, but the only information as to the department that a student writer comes from is in the sub-corpora of texts from that department. The smallest sub-corpus consists of texts from students of the Optometry and Radiography Department and contains about 4,200 words in three essays, while the largest is from students of the Institute of Textiles and Clothing, and contains about 99,000 words in 99 essays. Sinclair acknowledges the difficulty of attaining balance, stating that 'The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.'

Sinclair recommends that ‘Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.’ Cheng (2012, p. 31) lists a number of external criteria, including mode, which for both BAWE and PLEC is written; type of text, which in this study is academic essays; the domain of the text, which in this case is academic; the language varieties of the corpus, which is British English, although in native and non-native speaker variety; the location of the texts, which is England and Hong Kong; and finally the date the texts were written which is the early 21st century.

The final recommendation in Sinclair’s list is that ‘A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.’ For homogeneity he suggests that corpus compilers should ‘reject obviously odd or unusual texts.’ He defines rogue texts as those which ‘stand out as radically different from the others in their putative category, and therefore unrepresentative of the variety on intuitive grounds.’ It is assumed that the compilers of BAWE and PLEC will have been aware of this issue, an in this author’s examination of the texts during this study, his intuition has not been piqued by any possible rogue texts, and a visual inspection of the low-graded PLEC texts did not reveal any. However, there was only one A+ grade essay in PLEC (included in Appendix 5), and therefore statistics for this text may be outliers. Despite this, it was felt important to retain the essay in order to track proficiency level-related issues fully. In cases where the results for this essay are unexpected, this is mentioned in the following chapters.

3.2.2 The Corpus-based Approach

The essential characteristics of the corpus-based approach are, according to Biber, Conrad and Reppen (1998), that:

> - it is empirical, analyzing the actual patterns of use in natural texts;
> - it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
> - it makes extensive use of computers for analysis, using both automatic and interactive techniques;
> - it depends on both quantitative and qualitative analytical techniques. (p. 4).

The qualitative aspects of the approach are emphasized by Biber, Conrad and Reppen, who state that analyses must go beyond counts of linguistic features, and interpretation the patterns found is a vital part of the approach.

Uses of the corpus-based approach are identified by Biber, Conrad and Reppen (1998), and include the ability to focus on a linguistic feature, the identification of patterns of association of lexical items, and to focus on the characteristics of texts. In the operationalisation of this approach it is possible to identify features which may be lexical, grammatical, or register-specific. This enables the recognition of general patterns, for example in learners' interlanguage.

There are a number of limitations to corpus-based research, which are described by Szudarski (2018, pp. 9-10). Firstly, a corpus can show us only what it contains, and in the case of this research, the PLEC corpus does not contain texts on a range of disciplinary topics in the way BAWE and CJA14 do. This means that genre-based analysis, as recommended by Hyland (2008b) and Gardner and Nesi (2012), is much more restricted, even though the essays have been sorted into broad discipline categories, as their topics are not discipline-related. Also, because the essays were written for an English subject, students may have been following the teaching of that subject, and writing how their English teacher had taught them to write, rather than how they wrote in their disciplinary subjects.

The second limitation is that a corpus may be too small. In this research there is only one A+ grade PLEC essay, and 3 F grade PLEC essays, as graded by teachers for the assignment used to create the corpus. This means that it is difficult to compare high and low scoring essays to BAWE, which only contains high-scoring ones.

Thirdly, a corpus may present language out of its context. This limitation may be partially overcome by annotation or by the researcher's knowledge of the corpus collection and students. However, not all useful details may have been annotated. For example, the researcher knows the type of learner training that the PLEC students undertook, but this is not the case for BAWE, and the corpus, although annotated with many useful features, does not include this.

Some aspects of corpus-based research has been criticised by Tognini-Bonelli (2001). Firstly she states that it does ‘not leave the methodological and theoretical space’ to discover uses for language items outside the theoretical focus of the research (p. 66). To address this concern, in this thesis the corpus-based theory is used as a starting point, and not seen as a barrier to further investigation. An example of this is the section on word lists in Chapter 4.4 on Vocabulary, in which the research on the use of the Academic Word List leads to a wider discussion of the ‘common core’ hypothesis, and experimentation with other lists and corpora.

Secondly she states that corpus-based linguists view the relationship between theory and data as one in which the corpus evidence is seen as an extra bonus, rather than as a determining factor in analysis, and therefore cannot challenge theories. The approach to this concern in this thesis is that the author is not promoting a theory, and the theories are being tested to ascertain if their claims are supported by the data from the comparison of particular corpora. There are a number of cases in which the theories are not supported, for example in the use of 'existential *there*'. A table of the theories and whether the evidence supports them or not can be found in the summary of the research in Chapter Seven.

Thirdly she criticises the tagging of corpus data, and is concerned that such tagging follows pre-existing theories upon which the tagset is built, and that the corpus may then be analysed by the tagset, and the researcher will 'easily lose sight of the contextual features associated with a certain item and will accept single, uni-functional items – tags – as the primary data' (p. 73). Although this approach is used in this thesis, for example in the section on parts-of-speech, in which the keyness of various word classes is measured, this is used as a guide to further investigation, as can be seen in Table 4.3.1.1, rather than as an end in itself.

Her final criticism in her chapter on the corpus-based approach is that it will 'prioritise the information yielded by syntactic rather than lexical patterns' (p. 81). In this thesis lexical patterns are heavily represented, for example in recurrent word combinations, n-grams, conc-grams, phraseology, stance and voice.

Thus Tognini-Bonelli's concerns have been taken into account, and the negative effects of the prioritisation of theory have, it is hoped, been avoided.

3.2.3 The Corpus-driven Approach

Tognini-Bonelli (2001) then goes on to contrast corpus-based research with corpus-driven research. The differences are summarised in Cheng (2012, p. 187) who states that while the corpus-based approach is deductive, top-down, and proceeds from theory to hypothesis, observation, and then confirmation, the corpus-driven approach is inductive, bottom-up, and proceeds through stages of observation, pattern, tentative hypothesis, and ends with theory. Cheng notes how this has enabled the identification of language features that had previously not been known, and therefore not been included, in grammar books and dictionaries (pp. 188-9).

Despite the undoubted usefulness of the corpus-driven approach, for example in data-driven language learning activities in which students examine concordance lines and look for patterns of use, as reviewed in Chapter Six below, it is not used as the main approach for this thesis. The first reason for this is that the literature of CIA provides a broad and substantial theoretical foundation for corpus comparison, one which can be built on and has a research gap that can be addressed by this thesis. Secondly, the aim of this thesis is to inform language teaching in a variety of areas, and thus theories are needed that cover these areas. Corpus-driven research is by nature unpredictable, in that the patterns in the data are unknown before the analysis, and therefore it cannot be known in advance whether the theories that are derived from the data will inform language teaching in a variety of areas, or would instead be more fine-grained, such as Tognini-Bonelli's findings about the usage of *any* (2001, pp. 65-6).

Some aspects of the corpus-driven approach have also been criticised by McEnery, Xiao and Tono (2006, pp. 8-11) on a number of grounds. Firstly, they suggest that it is an unwarranted assumption that a large-enough corpus is representative due to its size. In this research PLEC-EAP and BAWE-EON are about 600,000 words each, far less than some reference corpora, and thus representativeness, which is the extent to which a sample includes the full range of variability in a population (Biber, 1993, p. 243), could be an issue. Secondly, McEnery, Xiao and Tono take issue with corpus-driven approach's rejection of tagging, regarding the necessary lack of preconceptions for principled corpus-driven tagging with corpus-derived tagsets as unrealistic, and they regard the approach as an 'idealized extreme' (p. 8). For all these reasons it was decided that a corpus-based approach best suits this thesis.

3.3 Research Questions

The research gap in Chapter One, combined with the previous research in Chapter Two and the research approach above, lead to the following research questions:

> To what extent are the commonly-taught aspects of academic essay writing and findings from the research literature on academic writing reflected in differences between the high standard essays by English native speakers in the BAWE corpus as compared to those of the Hong Kong students in the PLEC corpus?
>
> What changes would these differences (if any) suggest to the inclusion of these commonly-taught aspects?

To answer the first question requires an examination of existing teaching materials to ascertain what is already being taught, followed by a comparison between the corpora to see if there are skills that are not being taught that should be, or could be emphasised to a greater or lesser extent. The findings regarding the differences will have implications for teaching and learning that are addressed by the second question.

The objective of these questions is to lead to recommendations for new possibilities for input into academic writing courses, with the objective of improving the students' academic essay writing. The research methods by which this could be implemented are detailed below.

3.4 Research Methods: Strategy and Design

In outline, the research strategy in this thesis follows Cheng's (2012, p. 187) outline of the stages of a corpus-based approach: theory, hypothesis, observation and confirmation. The theory used is comparative interlanguage analysis, the hypothesis is that comparison of corpus language use can reveal patterns useful to learner writers, the observations form the study findings, and the confirmation is seen in the results of the study.

3.4.1 Procedure

The first stage in the research design is a survey of what is commonly taught in academic writing courses. This was done by a survey of subject materials from a number of Hong Kong universities and published academic writing books, including books intended to be course books as well as books for self-access study.

The results of this survey form a list of commonly-taught topics in academic writing. Based on the results of this survey, and on a review of the associated literature, the BAWE-EON and PLEC corpora were analysed to compare any differences in the realisation of these topics. To provide a more equitable comparison, a sub-corpus of the BAWE was extracted: the level one essays written by native English speakers called BAWE-EON. Essays were selected as the PLEC corpus is composed of entirely of essays, and native English speakers were isolated as this is the target language group for comparison. Despite the existence of research into Hong Kong English as a variety of English (Setter, Wong and Chan, 2010), and also into English as a lingua franca (Seidlhofer, 2011), a native speaker group were chosen because the official English variety of the university at which the PLEC students were studying is British English, and the BAWE native speakers are the closest group of students to this. The BAWE-EON sub-corpus includes 330 essays on a variety of subjects. As a part of this research a number of sub-sub-corpora were isolated from this, for example level one archaeology essays by native English speakers. This allows a limited comparison to the PLEC department-specific sub-corpora, but the departments are not exactly the same, so comparisons are limited in this respect.

3.4.2 Readability

Readability was measured using a number of methods: Flesch Kincaid Grade Level, Flesch Reading Ease, Gunning Fog Readability Index, Simple Measure of Gobbledygook (SMOG), and Coleman-Liau Readability Indexes, giving results measured in a number of years in the United States' education system.

3.4.3 Word Usage

Word usage was investigated through concordancer software, which on the input of a search term, output lists of lines containing those search terms in Key Word in Context (KWIC) format, which contain a key word, called a 'node' in the centre of each of multiple lines of text that form the context, and these lines can be analysed and counted.

3.4.4 Word Frequencies

Word frequencies were measured by the number of occurrences found in the corpora, and compared by normalising them as either instances per hundred thousand words or instances as percentages of the total number of words in the source corpus. Salience when comparing use of language was measured using Ahmad's (2005) weirdness formula, which divides the proportional use in the special corpus (in this case PLEC) by the proportional use in the general corpus (in this case BAWE-EON).

The formula for weirdness is:

$$\text{weirdness (term)} \quad = \quad \frac{f_{special} \, / N_{special}}{f_{general} \, / N_{general}}$$

where f = frequency, n = total number, special = the special language corpus and *general* means a general language corpus. The threshold for weirdness is 1.0 Also used to determine whether a comparison is significantly different enough so that it cannot be fulfilling the null hypothesis is a measure called ‘log likelihood’ and the related ‘p-level’ or probability level of the occurrence of an instance, where p = 0.01 is a 1% probability of getting a result when actually the result was unrelated. Log-likelihood has the advantage of not assuming that the data are normally distributed (McEnery, Xiao and Tono, 2006, p. 55). Effect size is measured using Log-ratio (Hardie, 2014).

3.4.5 Software

A variety of computer software was used in this research. Concordancing was carried out using Wordsmith 7 (Scott, 2017).) and concordancing using regular expressions to search for text utilised TextSTAT 2.9 (Hünning, 2000) and AntConc 3.4.4 (Anthony, 2014). Word frequency lists and n-grams were derived using ConcGram (Greaves, 2009) and analysed using Excel. The results of the survey of topics in academic writing courses were also tabulated using Excel. Readability statistics were analysed using two programs, Flesch 2.0 (Frink, 2007) and RocketReader Readability (Ronald, 2013). Grammatical accuracy was examined using LanguageTool 3.4 (Naber, 2016). Software used for more limited purposes is described in Chapters 4 and 5 below.

Based on this analysis using the above procedures and tools, suggestions are made for the inclusion of topics that should form the content of academic writing subjects for the Hong Kong students.

3.6 Summary

This chapter has demonstrated that the current research follows a theoretically-appropriate research method by explaining the conceptual framework of Contrastive Interlanguage Analysis on which the research is based, and reviewed criticisms of the framework that have led to its refinement. The way which the current research fits into this framework has been detailed, such as the comparison between PLEC as the learner corpus and BAWE-EON as the control corpus, followed by analysis using methods such as frequency counts and concordancing. To demonstrate proper data collection, Sinclair's (2004) corpus design principles were described and related to the choice of corpus and sub-corpus design.

Based on this framework, the research questions were given regarding whether and how aspects of academic writing differed from BAWE-EON to PLEC, and what changes these differences might suggest to the commonly-taught aspects.

The research method follows a well-established corpus-based approach of theory, hypothesis, observation and confirmation, based on a review of what is taught regarding academic essay writing. The hypothesis is that comparisons of corpus language use using standard corpus linguistics methods such as word usage and word frequency analysis can reveal patterns useful to learner writers. The observations form the study findings and confirmation, which are detailed in the next two chapters.

Chapter Four: Findings on Existing Materials, Grammar and Vocabulary

This chapter presents the findings of the study on grammar and vocabulary, and the implications of each finding. In accordance with the corpus-based methodology, the hypotheses from the research in the literature review and the methods from the conceptual framework are tested by comparing findings from the two main corpora, with associated findings from other corpora when relevant. These findings are then examined and possible reasons for the similarities and differences between the content of the corpora are given. However, the implications for teaching and learning and methods of implementing the results into language learning activities are discussed in more detail in the following chapters.

The chapter starts by examining existing EAP teaching materials to investigate what is commonly taught to students of academic essay writing, and to provide an organisational structure for the detailed research. After this, grammatical and lexical findings from the literature are compared to data from the corpora.

The analysis is conducted in a number of stages. Firstly existing teaching materials are reviewed in order to analyse what is currently regarded as valuable to teach in academic writing. The features revealed are then used to analyse the corpora, leading to a number of significant differences found between the BAWE and PLEC essays, and these are then discussed in relation to the concepts and theories from the literature review.

4.1 Existing Teaching Materials

The first stage of the study was to analyse existing academic essay writing input materials to investigate what is currently seen as valuable to teach students. These materials were two books from each of two Hong Kong universities, with one being more advanced. Five published books were also analysed:

1. Introduction to Academic Writing (Oshima and Hague, 2007)
2. Study Writing (Hamp-Lyons and Heasley, 2006)
3. Pathways: Writing Scenarios (McWhorter, 2010)
4. Student's Book of College English (Skwire, 2012)
5. The Academic Writer's Handbook (Rosen, 2012).

In each book the topics covered were analysed for frequency. Table 4.1 shows the results for features that occur in more than one book:

Table 4.1: Academic writing features and their frequency

Frequency	Features
7	Citations
6	Editing, Essay introduction moves, Quotations, Researching, Summary
5	Argumentative, Definitions, Essay Conclusions, Hedging, Paraphrase
4	Clauses in sentence structure, Comparison and contrast paragraphs, Paragraph structure, Peer reviews, Plagiarism, Planning, Punctuation, Run-on sentences, Sentence fragments, Synthesising, Thesis statement
3	Clustering, Coherence, Comma usage, Complex sentence structure, Reporting verbs, Subject-verb agreement, Supporting evidence sentences, Topic sentences
2	Adjectives and adverbs for describing, Analysis, Capitalization, Cause and effect essays, Compound sentences structure, Describing processes, Descriptive paragraphs, Essay body paragraphs, Essay organisation, Evaluation, Modifiers, Nominalisation, Note-taking, Organisation, Parts of speech / Word forms, Simple sentence structure, Subordinate clauses, Synonyms, Time order words, Transition signals, Writing process

Of these features a number were selected for comparison between the corpora, while for others no analysis was undertaken as it was not possible due to lack of evidence in the corpora: the latter features included the writing process, researching, note-taking, planning, editing, and peer reviews, which appeared in two, six, two, six, four and four texts respectively.

4.1.1 Analysis

The features were then grouped into areas for analysis. These areas followed the standard marking system used at the university where the PLEC corpus was collected, comprising major categories of content, organisation, language and conventions. The language category was divided into grammar and vocabulary. Conventions consisted of referencing, register and genre, as other sub-categories of format and layout from the university system did not appear as features in the instructional texts.

Of these categories, the one with features mentioned most frequently in texts was language with 55 mentions, consisting of 41 for grammar, 9 for vocabulary and 5 mentions for style and tone. This was followed by content, with 46 mentions in more than one text. Organisational features had 30 mentions; and finally conventions had 17. These categories were then investigated by analyses of the corpora. The following sections are therefore ordered in this way: language, including grammar and vocabulary, and in the following chapter, content, organisation and conventions.

Academic features from the teaching materials that are examined below include citations, editing, parts of speech, sentence structure, subordinate clauses, hedging, transition signals, processes, comparison and contrast connectors, reporting verbs, adjectives, adverbs, synonyms and nominalisation. The first section is on comparing language features, and starts with grammar and parts of speech.

4.2 Comparison of Language Features

Although it has been argued that there is a continuum from grammar to vocabulary, and these areas should be combined as lexicogrammar (Sinclair 2003, p. 63; Bennett, 2010, p. 10), this section organises them separately following the categorisation of the existing teaching materials described above. Despite this, there is a considerable degree of overlap between the areas, and for some research, for example on fixed multi-word constructions, there is a considerable lexical component as well as the grammatical organisation of the phrases.

4.3 Grammar

The difficulties that students have with grammar in academic writing were highlighted by Evans and Green (2007), who state that 'grammatical resources are also generally perceived as inadequate to meet the challenges placed on them in the production of academic assignments' (p. 14).

Five main areas concerning the use of grammatical features were analysed, and are presented in order of salience. They are: parts of speech (including articles, prepositions, determiners, and pronouns); and common errors, both of which supported the findings in the literature. Also analysed were tenses; syntactic complexity; grammatical slips and errors (which occur in both the expert and learner corpora). Finally there were fixed multi-word constructions, in which grammatically-correct word order is important. All but one of these analyses of features generated findings that can be applied to teaching and learning.

4.3.1 Parts of Speech

To compare the use of parts of speech in the PLEC and BAWE-EON corpora, the corpus files were tagged using TagAnt 1.2.0 (Anthony, 2015), and the tags were compared for keyness, which measures whether words 'are either unique to, or are found significantly more frequently in, a specialised corpus compared to a general reference corpus' (Cheng, 2012, p. 70). In this case the specialised corpus is PLEC-EAP and the reference corpus in BAWE-EON. This is shown in the table below, along with the section of this thesis in which the part of speech is analysed.

Table 4.3.1.1: Key part-of-speech tags in PLEC-EAP and BAWE-EON

Rank	Frequency	Keyness	Tag	Tag meaning	Thesis Section
1	32155	5343.961	VV	Verb, base form e.g. take	4.3.2 Tenses
2	19493	4449.117	MD	Modal	4.4.7 Repertoire of Recurrent Word Combinations
3	55627	2215.554	NNS	Noun plural	4.3.1 Parts of Speech
4	10381	2194.751	VVP	Verb have, present non-3rd person i.e. have	4.3.2 Tenses
5	38758	2167.428	SENT	End punctuation e.g. ?, !, .	4.4.9 Readability
6	15792	771.418	VVG	Verb, gerund/participle e.g. taking	4.3.2 Tenses
7	5992	574.408	VBP	Verb be, present non-3rd person e.g. am/are	4.3.2 Tenses
8	2156	509.573	I	First person pronoun	5.1.3 Disciplinary Variation
9	2570	482.967	VHP	Verb have, present non-3rd person e.g. have	4.3.2 Tenses
10	3959	446.828	JJR	Adjective, comparative	4.3.1 Parts of Speech
11	28112	230.833	PP	Personal pronoun	4.3.1 Parts of Speech
12	21489	215.691	TO	To	
13	114614	194.841	NN	Noun, singular or mass	See nominalisation in 5.3.4 Dimensions of linguistic variance
14	12530	183.654	VBZ	Verb be, present 3rd person singular i.e. is	4.3.2 Tenses
15	2086	158.838	RP	Particle e.g. give *up*.	4.3.1 Parts of Speech
16	545	125.670	VHG	Verb have, gerund/participle i.e. having	4.3.1 Parts of Speech
17	1253	118.305	JJS	Adjective, superlative	4.3.1 Parts of Speech
18	2265	101.927	EX	Existential *there*	4.4.7 Repertoire of Recurrent Word Combinations
19	1814	94.128	VH	Verb have, base form i.e. have	4.3.1 Parts of Speech

Spelling and grammatical errors in student texts could affect the accuracy of part-of-speech tagging, especially in the learner corpus. For example, when the 'F' grade PLEC essays were tagged with TagAnt (Anthony, 2015), the following tagging errors were found in the first essay (see Appendix Five for the full text of the essay):

- hong_NN kong_NNS

  Analysis: 'kong' should not be plural: it would be more correct to tag this hong_NP kong_NP where NP = 'Proper Noun'. Capitalising 'Hong Kong' fixes the problem.

- plastic_JJ waste_NN (correct), plastc_NN waste_NN

  Analysis: the spelling error of the missing 'i' in 'plastc' causes mis-tagging. Correcting the spelling solves the problem.

- It_PP will_MD suitiable_NN for_IN hong_NN kong_NNS ._SENT

  Analysis: the spelling error of the 'i' in 'suitiable' and/or the lack of 'be' after 'will' leads to 'suitable' being tagged as a noun instead of an adjective. Correcting the spelling solves the problem.

UCREL's WWW CLAWS tagger with the c7 tagset tagged all the previous examples correctly except 'hong kong', which it tagged hong_NN1 kong_NN1 (NN1 = singular common noun; e.g. book, girl.) Capitalising 'Hong Kong' solves the problem and tags it as NP1.

However, WWW CLAWS generates the following tagging errors:

- Recycling_NN1 is_VBZ a_AT1 good_JJ method_NN1 to_TO reduce_VVI losing_VVG matedal_JJ

  Analysis: 'matedal' probably means 'material' and should be tagged as a noun. The 'al' at the end of the word may cause the mistaken adjective tag.

- There_EX are_VBR recycle_VV0 policy_NN1 is_VBZ use_NN1 the_AT different_JJ colour_NN1 box_NN1

  Analysis: ‘use’ should be tagged as a verb because ‘is use’ a is grammatical error meaning ‘uses’.
- After_II that_DD1 ,_, the_AT separate_JJ waste_NN1 will_VM send_VVI to_II product_NN1 the_AT reuse_NN1 produce_VV0 ._.

  Analysis: ‘produce’ should be a tagged as an uncountable noun.
- We_PPIS2 can_VM know_VVI those_DD2 method_NN1 was_VBDZ not_XX a_AT1 good_JJ method_NN1 to_II used_JJ ._.

  Analysis: ‘used’ should be tagged as a verb, not an adjective.

However, the extent of these errors was tested by identifying those in the F grade PLEC essays, and less than two percent of the words contained spelling errors. Only four mis-taggings due to grammatical errors were found in the first ‘F’ grade PLEC essay. As shown by the high p levels in many of the findings in this research, the effects of these errors should be insignificant.

Parts of speech that were analysed included articles, prepositions, determiners, and pronouns, because Milton (2001, p. 11) identifies them as areas of concern in Hong Kong students’ interlanguage, with articles being avoided as an error-reduction strategy (p. 55). The tagged versions of the PLEC and BAWE-EON corpus were searched for the appropriate tagged words. As can be seen in Table 4.3.1.2 below, the figures show that the raw and normalised frequencies of articles, prepositions and determiners are higher in BAWE-EON than in PLEC, and the log-likelihood calculation shows that they are statistically significant. The normalised frequency is calculated per hundred thousand words in the tables in this thesis because the corpus sizes are in both around six hundred thousand words, so normalising to a million words would be overstating the occurrences. One hundred thousand words also usually gives frequencies as easy-to-read integers, as lower frequency vocabulary is sometimes normalised to just a few words.

Table 4.3.1.2: A comparison of words tagged as articles, prepositions, determiners, and pronouns in PLEC and BAWE-EON

Part of Speech	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p-level	PLEC use compared to BAWE-EON	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
Articles	48438	7568	57393	8731	1016.72	0.00	***	Under	-0.28
Prepositions	79641	12444	91116	13861	1108.68	0.00	***	Under	-0.23
Determiners	65319	10206	76229	11597	1158.41	0.00	***	Under	-0.26
Pronouns	29473	4605	27381	4165	31.26	0.00	***	Over	0.07

Note on p level: *** denotes p <.00001

The results in Table 4.3.1.2 above agree with Milton’s (2001) results, which found that in Hong Kong interlanguage, articles are under-used, first and second person pronouns are over-used, and many, but not all, prepositions are under-used. Despite this, the use of pronouns in PLEC is higher than those of BAWE-EON, which supports Li’s (2014) finding that Chinese speakers tend to use more pronouns than native speakers. Li concluded that sense of belonging is culturally important in Chinese and may be a cause of difference (p. 319). Word frequency comparison of the pronouns in Table 4.3.1.3 below shows that the personal pronoun *my* is of much higher frequency in PLEC (0.0773) than BAWE-EON (0.0370), which follows Li’s interpretation.

Table 4.3.1.3: Top 12 under- and over-used pronouns in PLEC-EAP and BAWE-EON

	Pronoun	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p-level	Use in PLEC	Log ratio
		Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
1	some	3139	477.79	903	141.09	1250.31	0.00	***	Over	1.76
2	I	2178	331.51	1010	157.81	407.42	0.00	***	Over	1.07
3	many	2160	328.77	1042	162.81	369.41	0.00	***	Over	1.01
4	more	3209	488.44	1866	291.56	324.70	0.00	***	Over	0.74
5	it	7837	1192.87	5777	902.64	260.31	0.00	***	Over	0.40
6	them	1848	281.28	997	155.78	236.27	0.00	***	Over	0.85
7	enough	379	57.69	137	21.41	111.70	0.00	***	Over	1.43
8	**my**	508	77.32	237	37.03	93.77	0.00	***	Over	1.06
9	most	1271	193.46	833	130.15	80.53	0.00	***	Over	0.57
10	everyone	175	26.64	53	8.28	65.60	0.00	***	Over	1.68
11	you	422	64.23	315	49.22	12.86	0.00	***	Over	0.38
12	those	702	106.85	601	93.90	5.37	0.00	**	Over	0.19
1	his	117	17.81	2378	371.55	2575.21	0.00	***	Under	-4.38
2	her	83	12.63	1120	175.00	1091.65	0.00	***	Under	-3.79
3	him	14	2.13	489	76.40	582.19	0.00	***	Under	-5.16
4	what	326	49.62	999	156.09	376.51	0.00	***	Under	-1.65
5	which	1661	252.82	2841	443.90	345.25	0.00	***	Under	-0.81
6	himself	9	1.37	229	35.78	259.25	0.00	***	Under	-4.71
7	itself	31	4.72	289	45.16	246.94	0.00	***	Under	-3.26
8	this	3951	601.38	5434	849.05	276.61	0.00	***	Under	-0.50
9	one	1137	173.06	1748	273.12	147.22	0.00	***	Under	-0.66
10	each	164	24.96	403	62.97	110.45	0.00	***	Under	-1.34
11	such	1005	152.97	1520	237.50	119.99	0.00	***	Under	-0.64
12	any	323	49.16	621	97.03	103.83	0.00	***	Under	-0.98

Note: Sig. = Significance; p level: *** denotes p <.00001

Hong Kong students often use the phrase *in my (personal) opinion*, so this was searched for, and 12 occurrences were found (1.87 per hundred thousand) compared to 8 (1.22 per hundred thousand) in BAWE-EON, so it seems that this expression could be one cause of the difference, and that Hong Kong students could be advised to use this expression sparingly.

Despite this, removing the occurrences of *in my (personal) opinion* from the figures still gives a normalised frequency for *my* of 77 in PLEC and 34 in BAWE-EON, a significant difference, and showing that *my* is still more common in PLEC, so Li’s (2014) conclusion may still be correct.

The importance of sense of belonging may be reflected in the greater importance of collective pronouns. Leedham (2011, p. 236), based on an analysis of the BAWE corpus, found that Chinese students make greater use of first person plural 'We'. Regarding the first person plural, in the PLEC corpus 1,220 instances of *we* were found, or 0.1856 %. In the BAWE-EON essays the total was 1,393, or 0.2177%, which is more than the percentage from PLEC. The sum of the normalised frequencies for the three first-person plural pronouns including possessives, *we, us* and *ours*, was greater in the BAWE-EON at 289, in comparison to 245 in PLEC thus not supporting Leedham's assumption that it is the language background of the students that is significant in the case of PLEC. This may be because there is some evidence that students who represent themselves as part of the research community by 'interacting in an overtly cognizant and intertextual manner with readings relevant to their topics' achieve higher scores (Swales 2014, p. 138), and the BAWE essays are all high scoring essays, but the PLEC essays are not.

The ratio of these grammar words (rather than content words such as nouns), is discussed further in the light of the readability findings below.

Similar in appearance to prepositions are particles, which will be defined here as words that look like adverbs or prepositions in a phrasal verb, for instance the *up* in *give up*. Milton (2001, p. 72) found that in his corpora that Hong Kong students used about 30% of the number of phrasal verbs that UK students used. Particles were identified in the keyness calculation as key in the comparison between PLEC-EAP and BAWE-EON. A total of 2,086 were found in PLEC-EAP compared to 1,366 in BAWE-EON, giving a log-likelihood of 171.13 and a significance of 0.000, so they are greatly over-used in PLEC-EAP. To address this, students can be recommended to use more formal single-word verbs in academic essays, for example *discover* instead of *find out*.

Nouns, both singular and plural, were identified by the keyness statistics as different in the two corpora. Milton (2001) comments on the over-use of plural nouns in his corpora of Hong Kong interlanguage. He points out that this is unexpected from a contrastive analysis perspective, as Cantonese does not mark plural nouns. In addition, he claims that ‘plural nouns are rejected as sentence subjects by Chinese syntax’ (p. 18). He found that HK students used only about 1% less plural common nouns (9.37%) as singular nouns (10.28%).

The use of plural nouns, tagged as NNS, and singular and uncountable nouns, tagged as NN, in the BAWE-D and PLEC-EONS corpora were analysed, as can be seen in Table 4.3.1.4 below. BAWE-D was used to check if there was significant variability within disciplines, but there was nothing noticeable.

Examining these results and comparing them to Milton’s findings, there is over-use of all these nouns. However, contrary to his findings, PLEC-EAP students used over twice the percentage of singular and uncountable nouns compared to plural nouns, more similar to the proportion used by his UK students and the BAWE-D students.

Table 4.3.1.4: Noun uses in PLEC-EAP and BAWE-D

Part of speech	PLEC-EAP		BAWE-D		Log-likelihood	Sig.	p-level	Use in PLEC	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
All plural nouns	55627	8413	41747	6158	1631.87	0.000	***	Over	0.38
Sentence-initial plural nouns	966	146	705	104	34.26	0.000	***	Over	0.42
All singular & uncountable nouns	113681	17193	109376	16133	7.89	0.000	***	Over	0.02

p level: *** denotes p <.00001

The reasons for this could include that his figures are for all the students who took the Use of English exam, of whom only the better ones qualified for university. Given that the differences in the use of all types of these nouns between PLEC-EAP and BAWE-D is less than 2.5%, and that nominalisation is a feature of academic writing, as is discussed below, there seems little point in encouraging students to use less nouns.

Comparative adjectives were also found to be key in the comparison of parts of speech in PLEC-EAP and BAWE-EON. In PLEC-EAP, 53 different comparative adjectives were found, compared to 90 in BAWE-EON. The table below shows a comparison of comparative adjectives in PLEC-EAP and BAWE-EON. Adjectives that are not used in both corpora, and are not significantly different, have been removed from the table. Only one PLEC-EAP adjective with a raw frequency of greater than one and which was not used in BAWE-EON was *fresher*, with 7 occurrences. BAWE-EON adjectives which were not used in PLEC and had 7 or more occurrences were *broader* (25), and *darker* (7).

It can be seen in Table 4.3.1.5 below that common words such as *more*, *worse*, *better*, and *less*, which are on the General Service List (West, 1953) of the commonest headwords in English, top the table and are significantly over-used in PLEC-EAP, while more specific adjectives, such as *stronger* and *wealthier*, and the words not used in PLEC above, are under-used. This may be because the learner writers in PLEC have a more restricted vocabulary, as is discussed in the Vocabulary sections below.

Table 4.3.1.5: Comparative adjectives in PLEC-EAP and BAWE

Comparative adjective	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p-level	Use in PLEC	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
more	3209	501.40	1866	283.87	324.70	0.00	***	Over	0.74
lower	405	63.28	148	22.52	117.39	0.00	***	Over	1.41
worse	133	20.78	26	3.96	75.93	0.00	***	Over	2.32
better	371	57.97	180	27.38	62.60	0.00	***	Over	1.00
fewer	125	19.53	28	4.26	63.91	0.00	***	Over	2.12
less	547	85.47	316	48.07	56.58	0.00	***	Over	0.75
stricter	43	6.72	3	0.46	40.53	0.00	***	Over	3.80
cheaper	57	8.91	15	2.28	25.01	0.00	***	Over	1.89
safer	38	5.94	7	1.06	22.66	0.00	***	Over	2.40
further	151	23.59	324	49.29	69.19	0.00	***	Under	-1.14
wider	6	0.94	67	10.19	61.36	0.00	***	Under	-3.52
earlier	26	4.06	81	12.32	31.16	0.00	***	Under	-1.68
smaller	11	1.72	49	7.45	27.03	0.00	***	Under	-2.19
matter	82	12.81	153	23.28	23.73	0.00	***	Under	-0.94
greater	113	17.66	194	29.51	23.84	0.00	***	Under	-0.82
older	8	1.25	32	4.87	16.07	0.00	***	Under	-2.04
closer	18	2.81	45	6.85	12.69	0.00	***	Under	-1.36
stronger	8	1.25	27	4.11	11.41	0.00	**	Under	-1.79
wealthier	1	0.16	10	1.52	8.79	0.00	**	Under	-3.36

Note: p level: *** denotes p <.00001; ** denotes p < .0001

A search was conducted for comparatives starting with *more*; e.g. *more expensive*, however, the raw frequencies for each phrase were very low, at less than 5 and usually lower than 3 in both corpora, therefore they have not been included in the analysis.

Given the over-use of these general adjectives, students could be encouraged to use more specific adjectives in their work. The use of adjectives is further analysed and discussed in the section on disciplinary variation below.

A similar analysis was undertaken for superlative adjectives. The only superlative in PLEC-EAP that was not in BAWE-EON and had a normalised frequency of over one was *busiest*, with 25 occurrences. Superlatives with a normalised frequency of over one that were in BAWE-EON, but not in PLEC-EAP were *earliest (20), closest (10),* and *oldest (7).*

Statistically-significant use of superlative adjectives in PLEC-EAP and BAWE-EON are shown in Table 4.3.1.6 below. In Table 4.3.1.6 it can be seen that the superlatives which were significantly over-used in PLEC-EAP are general words about quantity and quality, while the under-used ones are about a variety of more specific extremes, a pattern which parallels the comparative adjectives. Another parallel is that superlative adjective phrases, such as *most valuable* all have very low frequencies in the corpora, as did the comparative phrases. Therefore the recommendation is the same as for the comparative adjectives: more specificity, and the use of more multi-syllabic adjectives that require *more* or *most* before them, such as *more expensive*.

Table 4.3.1.6: Superlative adjectives in PLEC-EAP and BAWE-EON

Superlative adjective	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p-level	Use in PLEC	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
most	1271	198.59	833	126.72	80.53	0.00	***	Over	0.57
best	236	36.87	154	23.43	15.25	0.00	***	Over	0.58
worst	31	4.84	11	1.67	9.39	0.00	**	Over	1.46
greatest	6	0.94	51	7.76	41.87	0.00	***	Under	-3.13
largest	3	0.47	36	5.48	33.80	0.00	***	Under	-3.62
highest	2	0.31	30	4.56	30.15	0.00	***	Under	-3.95
poorest	1	0.16	10	1.52	8.79	0.00	**	Under	-3.36
lowest	3	0.47	13	1.98	7.01	0.01	*	Under	-2.15
strongest	3	0.47	13	1.98	7.01	0.01	*	Under	-2.15
biggest	5	0.78	14	2.13	4.68	0.04	*	Under	-1.52

Note on p level: *** denotes p <.00001; ** denotes p < .001; * denotes p < .05

The use of *have* and *having* was also identified by the keyness analysis of the parts of speech was, tagged as VH and VHG respectively. These were compared for PLEC-EAP and BAWE-EON, and the results can be seen in Table 4.3.1.7 below.

Table 4.3.1.7: Use of 'have' and 'having' in PLEC-EAP and BAWE-EON

Part of speech	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p-level	Use in PLEC	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
VH (have)	1814	283.43	1299	197.61	72.39	0.000	***	Over	0.44
VHG (having)	545	85.15	241	36.66	112.72	0.000	***	Over	1.14

Note on p level: *** denotes p <.00001

As can be seen in the table above, there is over-use of both forms of *have*. Where appropriate, students could be encouraged to use more-specific synonyms of *have*, such as *possess* or *own*.

4.3.2 Tenses

A comparison of present tenses of different aspects was made by Hu and Gu (2015) between the Chinese Learner English Corpus (CLEC) and BAWE. They used sub-corpora consisting of the texts in the St3 sub-corpus of the CLEC, written by CET-4[3] level students and made up of 169,386 tokens, and a topically-parallel sub-corpus of undergraduates' written production on arts and humanities and social sciences from the BAWE corpora, comprising 368,303 tokens.

They found that in Chinese students' writing, simple present tense and present progressive tense are overused, while present perfect tense and present perfect progressive tense are underused, but only the statistics for the present perfect tense are significant, due partly to the low frequencies of 32 occurrences in BAWE and 20 in CLEC. Regarding the reasons, Hu and Gu comment on the different view of time in Chinese. They describe English conceptions of time as independently flowing entity, whereas in Chinese 'the observer travels with the time mentioned in his imagination' (p. 143). A further complication is that Li and Luk state that "Chinese (and Cantonese), unlike English, does not have tense markers; it only has aspect markers" (2017, p. 65).

Hu and Gu comment that simple present tense ranks as the most frequent tense 'because it is grammatically and semantically easier, and more familiar to Chinese learners' (p. 145). Reasons for the over-use of the present progressive are that it is taught before the perfect, and Chinese students may not know that progressive tenses are rare in academic writing.

[3] CET-4 is the College English Test, level 4, which is mandatory for university students who are not English majors and is necessary for university graduation in mainland China.

Regarding the present perfect tense, they comment that ‘Chinese students may consciously or subconsciously try to avoid the use of perfect tense’ (p. 145) as Chinese learners find it difficult because in Chinese, adverbials are used to indicate the perfect, rather than verb inflection. The present perfect progressive contains not only the perfect, but also the progressive aspect, and therefore ‘is both grammatically and semantically challenging for Chinese learners’ (p. 145). This may result in an avoidance strategy by the students, who worry about making errors in these tenses.

To investigate whether this was true of the Hong Kong learners in the PLEC corpus, similar counts and statistical calculations were carried out, as can be seen in Table 4.3.2 below. A part-of-speech tagged version of the corpus was searched for tenses using regular expressions.[4]

Table 4.3.2: Comparison of present tenses in PLEC-EAP and BAWE-EON

Tense and aspect	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p-level	Use in PLEC	Log ratio
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000					
Simple present	15601	2359	14865	2193	3.56	0.00	***	Over	0.03
Present progressive	984	149	485	72	159.86	0.00	***	Over	0.98
Present perfect	1001	151	1198	177	23.33	0.00	***	Under	-0.30
Present perfect progressive	58	9	39	6	3.26	0.04	*	Over	0.53

Note on p level: *** denotes p <.00001; * denotes p < .05

[4] Tagged versions of the corpora were searched for simple present tenses by finding VVP (verb, present, non-3rd person) and VVZ (verb, present 3rd person singular) tags in the word lists. The perfect tenses were searched for using regular expressions: \w+_(VB[ZP]) +\w+_VVG for present progressive, \w+_VH[ZP] +(\w+_RB([RS])?)? *\w+_VVN for present perfect, and \w+_VBN +(RB(R|S)?)? *\w+_VVG for present perfect progressive. In these, \w+ stands for a word, and the upper case letters stand for the tags. The last two expressions contain code for possible adverbs in case writers had used any between the auxiliary and the main verb.

As can be seen in the table above, the result for the simple present and present progressive showed over-use, in accordance to Hu and Gu's findings. The present progressive was under-used as predicted, but the present perfect progressive was significantly over-used in PLEC. However, the raw frequencies are very low, because as Hu and Gu say, 'this tense is scarcely used in English writing' (p. 145).

The authors recommend more instruction in tenses for writing, as their interviews revealed that 'only a small fraction of the students reported receiving regular writing instruction', since writing and grammar are taught separately. One possible method of addressing this would be to teach the use of the present perfect from a functional perspective, such as for reporting correlations in literature reviews; e.g. *X has been found to increase with Y.* Although writing and grammar are taught separately in Hu and Gu's context, the PLEC-EAP students had gone through a different educational system in Hong Kong, so this might not be the case in their context, so instead the differences in the underlying structure of the Chinese and English temporal system may be the root cause of this situation.

4.3.3 Syntactic Complexity

One possible measure of the proficiency of students with academic writing is an assessment of the syntactical complexity of their essays. A computer program called the ‘L2 Syntactic Complexity Analyzer’ (L2SCA) has been developed and described by Lu (2010) and Lu and Ai (2015). They state that ‘some measures of syntactic complexity may be reliably used to differentiate levels of L2 proficiency’ (Lu and Ai, 2015, pp. 16-7) because it is a dynamic property of the learners’ interlanguage.

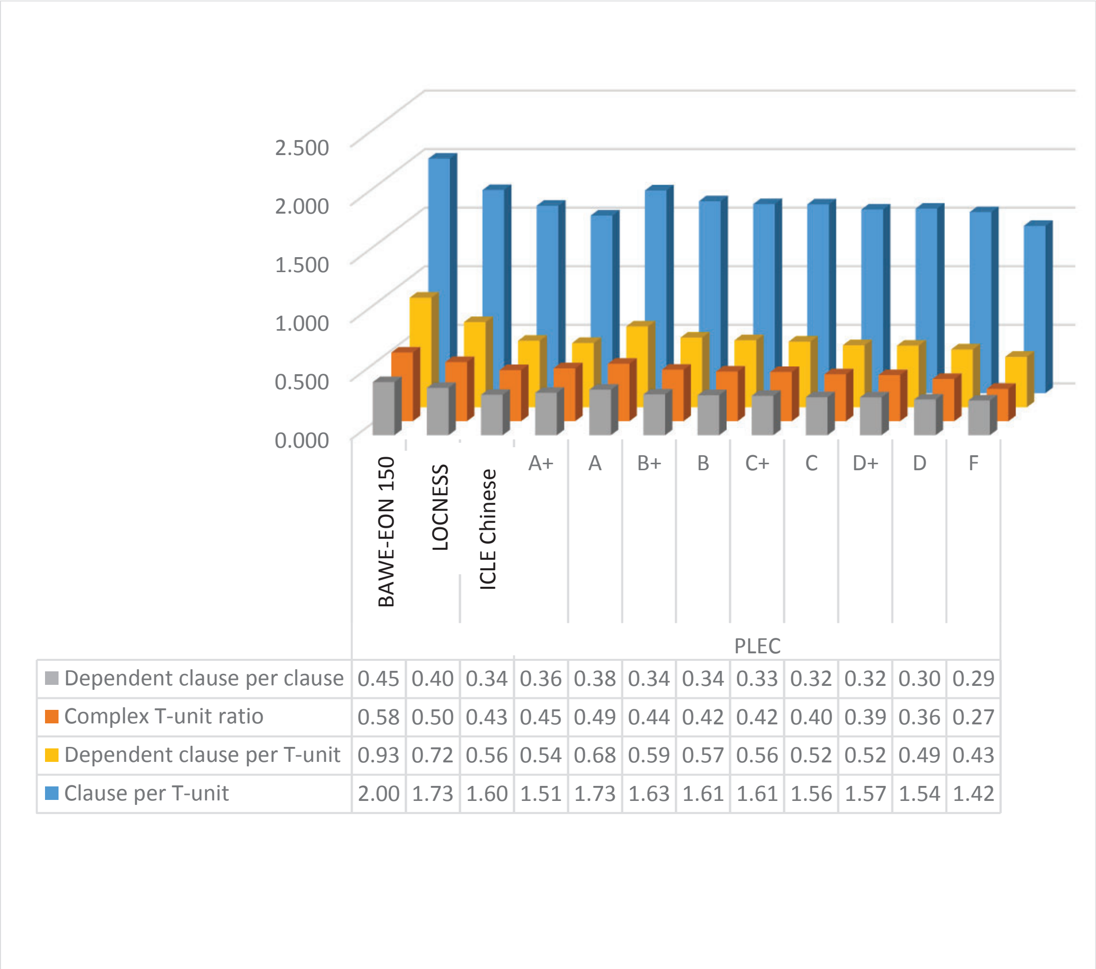
Regarding possible limitations of use of this program to compare the PLEC and BAWE-EON essays, relevant factors are that task complexity has not been found to influence syntactic complexity in L2 writing (Lu and Ai, 2015, p. 18), so there should not be an effect of the generic topics used in PLEC as compared to the discipline-specific ones used in BAWE. However, time-limited essay writing has been found to give rise to less complexity, and this may make the complexity of the BAWE essays higher because according to Leedham’s (2014) comparison of BAWE and ICLE, the BAWE students had ‘unlimited time for preparation and drafting’ (p. 3).

The L2SCA software uses 14 measures of syntactic complexity in four groups, which are based on mean counts of words, sentences, verb phrases, clauses, and ‘T-units’, which means ‘minimal terminable units’ and are a measure of syntactic complexity, each T-unit consisting of ‘one main clause with all the subordinate clauses attached to it’ (Hunt, 1965, p. 20). The first group of measures is length of production unit, which consists of mean length of clause, mean length of sentence and mean length of T-unit. The second group is amount of subordination, which is measured by clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, and dependent clauses per T-unit. The third group measures the amount of coordination in the text, and is assessed by the coordinate phrases per clause, coordinate phrases per T-unit, and T-units per sentence. The degree of phrasal sophistication is judged by complex nominals per clause, complex nominals per T-unit, and verb phrases per T-unit. Finally, overall sentence complexity is measured by the number of clauses per sentence. In addition to the computer rating, Lu (2010) tested their system by having human annotators give scores, resulting in correlations from 0.834 to 1.000.

The software was used by Lu and Ai (2015) to compare 200 argumentative essays from the LOCNESS corpus of British and American argumentative essays with 1,400 argumentative essays from the International Corpus of Learner English 2.0 (ICLE2.0), including essays by Chinese writers. While comparison of native versus non-native speaker writers only showed significant differences in three of the measures: mean length of clause, complex nominals per clause, and complex nominals per T-unit, when the L2 writers were compared against L2 writers from other language backgrounds there were significant differences in all 14 measures. This was probably because these groups had different proficiency profiles on the Common European Framework of Reference for Language (CEFRL), with Chinese learners having the highest number of B2 or lower ratings, and only one C1 rating, making them the lowest proficiency group in the comparison, and described as being at 'upper intermediate' level (Lu and Ai, 2015, p. 21).

The results of analysing the PLEC-EAP corpus, as can be seen in Figure 4.3.3 below, show the amount of subordination reflected the EAP grades from A to F, measured by clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, and dependent clauses per T-unit. The A+ grade is slightly anomalous, because there is only one essay in this grade. The amount of subordination in PLEC-EAP was lower than in ICLE 2.0 Chinese, which was lower than LOCNESS, which in turn was lower than in BAWE-EON. In the analysis shown in the figure below, only 150 essays were used for BAWE-EON, because the L2SCA software is limited to a maximum word length per essay of 2,000 words, and 150 BAWE-EON essays, or about 45%, were within this length restriction.

Figure 4.3.3: Comparison of Subordination Measures in PLEC-EAP, BAWE-EON, LOCNESS and ICLE 2.0 Chinese



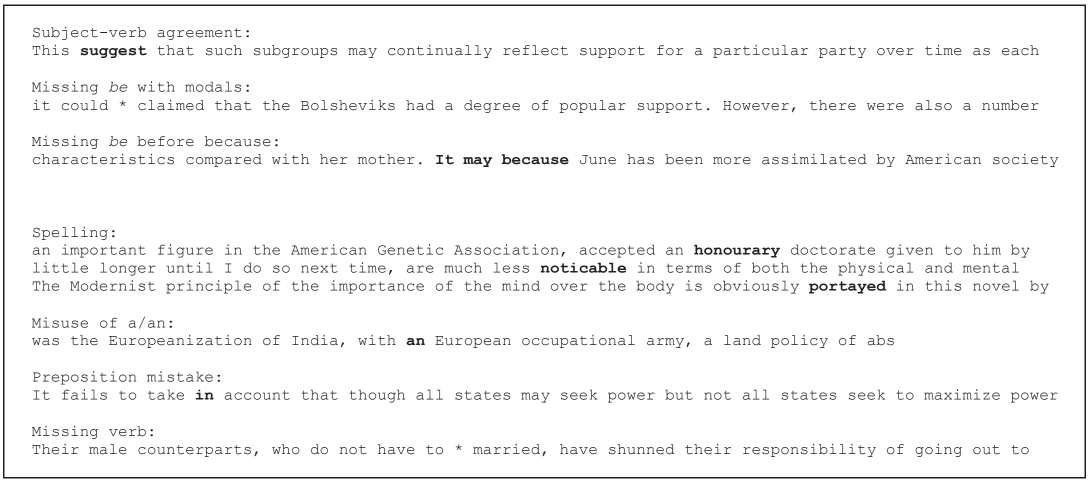
	BAWE-EON 150	LOCNESS	ICLE Chinese	A+	A	B+	B	C+	C	D+	D	F
■ Dependent clause per clause	0.45	0.40	0.34	0.36	0.38	0.34	0.34	0.33	0.32	0.32	0.30	0.29
■ Complex T-unit ratio	0.58	0.50	0.43	0.45	0.49	0.44	0.42	0.42	0.40	0.39	0.36	0.27
■ Dependent clause per T-unit	0.93	0.72	0.56	0.54	0.68	0.59	0.57	0.56	0.52	0.52	0.49	0.43
■ Clause per T-unit	2.00	1.73	1.60	1.51	1.73	1.63	1.61	1.61	1.56	1.57	1.54	1.42

In a comparison of the results for the different corpora, the mean length of sentence figures show that the sentence length in BAWE-EON is 26, 10 words or more longer than in PLEC-EAP A+ grade essays. Since the measures of syntactic complexity are all normed, this should not affect the results. Full statistics are given in Appendix One. The figures suggest that more attention could be paid to teaching subordination in sentence structure, which might also increase the mean sentence length to a figure closer to the BAWE-EON mean number of words per sentence.

4.3.4 Grammatical Errors

Evans and Morrison also cited graduating students as being "far from satisfied with their … imperfect mastery of grammar" (2012, pp. 40-1). Indeed the PLEC corpus contains a wide variety of grammatical errors. However, the BAWE corpus is far from error-free. For this thesis the BAWE-EON sub-corpus was scanned for a number of common errors made by Hong Kong students, and examples are shown in the figure below (omissions are marked with an *).

Figure 4.3.4: Concordances of grammatical errors in BAWE-EON



Subject-verb agreement:
This **suggest** that such subgroups may continually reflect support for a particular party over time as each

Missing *be* with modals:
it could * claimed that the Bolsheviks had a degree of popular support. However, there were also a number

Missing *be* before because:
characteristics compared with her mother. **It may because** June has been more assimilated by American society

Spelling:
an important figure in the American Genetic Association, accepted an **honourary** doctorate given to him by
little longer until I do so next time, are much less **noticable** in terms of both the physical and mental
The Modernist principle of the importance of the mind over the body is obviously **portayed** in this novel by

Misuse of a/an:
was the Europeanization of India, with **an** European occupational army, a land policy of abs

Preposition mistake:
It fails to take **in** account that though all states may seek power but not all states seek to maximize power

Missing verb:
Their male counterparts, who do not have to * married, have shunned their responsibility of going out to

Therefore it is recommended that, because even native speakers cannot avoid grammatical errors, as well as being taught accurate grammar and proof-reading strategies, students should be directed to proof-reading resources. These can be as simple as using the grammar checker built in to their word processors, or by using more specialist programs.[5]

[5] These include such as the online and downloadable grammar and style checker called Language Tool (https://www.languagetool.org) based on Naber's (2003) research, or a more student- and writing-task-specific tool such as this author's Common Error Detector webpage (http://www2.elc.polyu.edu.hk/cill/errordetector.htm) , which is designed to find, highlight, and suggest corrections for common errors made by Hong Kong students in academic essay writing. It also has functions specific to academic writing and which are not found in spell- and grammar-checkers, such as checking the format of references.

Students in some institutions may have access to automated feedback systems, for instance Educational Testing Service's 'Criteria' system (Burstein, Chodorow and Leacock, 2004), which provides feedback on word choice; grammar, usage and mechanics – conventions, organisation, development and style. The system was reviewed by Long (2013), who found that it was useful in providing feedback on some aspects of grammar, mechanics, usage, and style.

Computerised proof-reading may be the most efficient strategy, due to a number of factors in the Hong Kong tertiary academic writing setting. These factors include that only a limited number of errors may result in mis- or incomprehension on the part of faculty staff in the disciplines who are the readers of much of the students' writing on their degree programmes, as they are often Cantonese or Mandarin speakers who may understand mistakes caused by first-language interference errors. Secondly, some errors may involve aspects that could be said to be redundant, or at least non-intrusive, such as subject-verb agreement for third person final 's'. Therefore it may not be worth a student's time or effort to proof-read for these manually, unless the writing is to be graded on the grammar with sufficient rigour that reliance on computerised proof-reading alone is not enough. Priority can be given instead to vocabulary to communicate meaning accurately and in the language of the discipline, and to register in order to maintain an appropriate relationship with the reader.

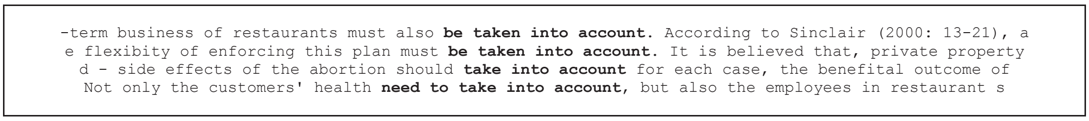
4.3.5 Fixed Multi-word Constructions

Consideration of grammatical correctness by students is necessary when they write some types of set phrases. Such phrases include ‘fixed’ and ‘unfixed’ multi-word constructions, which are of two types, according to Liu (2012, p. 33), who recommends that fixed multi-word constructions such as ‘in terms of’ may be learned in unanalysed chunks, whereas unfixed ones in which the word order differs or different words may be inserted into the construction, such as ‘take / be taken into account’, should be grammatically analysed so that students can use them accurately in production.

The BAWE-EON sub-corpus contains 154 instances of ‘in terms of’ in 330 essays, a percentage of 0.024% of the total words. The PLEC contains proportionately less at 48 instances in 8,000 scripts, or 0.0075 percent of the words.

The PLEC corpus contains four sentences that use ‘take / be taken into account’ (see Figure 3), of which half are incorrect.

Figure 4.3.5: Uses of ‘take / be taken into account’ in PLEC.



```
-term business of restaurants must also be taken into account. According to Sinclair (2000: 13-21), a
e flexibity of enforcing this plan must be taken into account. It is believed that, private property
 d - side effects of the abortion should take into account for each case, the benefital outcome of
 Not only the customers' health need to take into account, but also the employees in restaurant s
```

These are all from different essays, with the erroneous sentences from C and C+ graded essays. The corpus contains 38,680 sentences, giving one use of the term for every 9,670 sentences. With 8,000 essays in the sub-corpus, the term is used on average once about every twenty essays.

The BAWE-EON corpus contains 28 uses of the term, all grammatically correct. The sub-corpus contains 24,572 sentences, giving one use every 877 sentences. With 330 essays in the sub-corpus, the term is used on average once about every twelve essays, with a maximum of twice per essay according to AntConc's Concordance plot.

Although it seems that the term is used about twice as often in BAWE-EON, the overall rarity of the term raises questions about whether it is worth spending class time on.

Although Liu provides lists of the constructions, they are not categorised into fixed and unfixed, and they are lemmatised, so removing the grammatical variability inherent in unfixed constructions, and therefore making further research difficult. More analysis of phrases, including details of lemmatised lexical items and those in 'word families' can be found in the next section on Vocabulary.

4.4 Vocabulary

The importance of students' need for vocabulary was highlighted by Evans and Green (2007), who stated that 'inadequate receptive and productive vocabulary in English is the main problem' confronting the almost 5,000 students that they studied at the same university where PLEC was collected (p. 14). In order to improve this situation, the following section examines and analyses vocabulary teaching issues from the literature by applying the research findings to PLEC, BAWE-EON and CJA14.

The vocabulary features below are organised by length, from single words, to phrases, and finally to sentence length features. Single word features include key words, word lengths, type-token ratios, norm-specific vocabulary, and words from the Academic Word List (AWL) (Coxhead, 2000a). Multi-word phrases include lexical chunks from the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010), repertoire of recurrent word combinations, and n-grams. The sentence length feature concerns readability statistics. Findings for most of these were in accordance with the literature, but there were a number of differences. In addition, the corpus comparison showed that there were cases of over- and under-use, which in some cases were very significant.

4.4.1 Key words

A frequency list of key words in the PLEC-EAP corpus was generated using Wordsmith 7's KeyWords function and using BAWE-EON as the reference corpus. This list is to identify which words in PLEC-EAP occur with unusual frequency. Table 4.4.1.1 below shows the top twenty keywords. As can be seen in the table, the words are all related to the essay topics given to the PLEC students, for example topics on credit card use by students, smoking, and recycling.

Table 4.4.1.1 Top 20 Key words in PLEC-EAP as compared to BAWE-EON

	Key word	Freq.	Log Likelihood	Log Ratio	P	Lemmas
1	Hong	6,240	8,773.56	11.64	0	
2	Kong	5,934	8,375.14	142.21	0	
3	student	11,196	8,303.25	7.62	0	student[6136] students[5060]
4	card	7,930	7,494.31	8.54	0	card[5442] cards[2488]
5	credit	5,366	7,198.63	7.3	0	
6	students	5,060	6,917.17	8.09	0	
7	smoke	7,859	6,325.84	7.76	0	smoke[4659] smoking[3200]
8	smoking	3,200	4,384.56	8.22	0	
9	recycle	5,572	3,998.45	141.14	0	recycle[2833] recycled[257] recycling[2482]
10	restaurant	5,293	3,913.67	6.54	0	restaurant[2999] restaurants[2294]
11	recycling	2,482	3,503.05	140.95	0	
12	cards	2,488	3,404.54	8.15	0	
13	ban	4,107	3,269.85	8.26	0	ban[2385] banned[358] banning[1364]
14	waste	2,574	3,064.00	5.89	0	waste[2432] wastes[106] wasting[36]
15	restaurants	2,294	3,059.97	7.11	0	
16	abortion	2,167	2,793.87	8.22	0	abortion[2039] abortions[128]
17	cyber	1,775	2,505.20	140.47	0	
18	debt	2,363	2,337.74	6.17	0	debt[1825] debts[538]
19	cafe	2,908	2,297.45	9.14	0	cafe[1656] cafes[1252]
20	mainland	1,642	2,236.58	7.91	0	

This is a consequence of the different selection criteria for the corpus texts, in which BAWE students wrote on disciplinary topics, but PLEC students, in common with all ICLE corpus writers, wrote on general topics. The use of general topics is due to the need to devise topics that all students will have enough background information to write about with no research necessary in advance, which prevents students from memorising text and including it in their essays, although, as Milton (2001) points out, students can still memorise generic phrases that are suitable for any essay.

Table 4.1.1.2 Top 20 Lemmatised keywords in PLEC-EAP as compared to BAWE-EON

Rank	Key word	Freq.	Log Likelihood	P	Lemmas
1	student	11,196	8,303	0	student[6136] students[5060]
2	card	7,930	7,494	0	card[5442] cards[2488]
3	smoke	7,859	6,326	0	smoke[4659] smoking[3200]
4	recycle	5,572	3,998	0	recycle[2833] recycled[257] recycling[2482]
5	restaurant	5,293	3,914	0	restaurant[2999] restaurants[2294]
6	professional	4,389	2,210	0	professional[2428] professionals[1961]
7	ban	4,107	3,270	0	ban[2385] banned[358] banning[1364]
8	problem	3,724	1,528	0	problem[2549] problems[1175]
9	accord	3,629	1,658	0	accord[1815] according[1814]
10	advantage	2,916	1,640	0	advantage[1689] advantages[1227]
11	cafe	2,908	2,297	0	cafe[1656] cafes[1252]
12	smoker	2,705	1,986	0	smoker[1427] smokers[1278]
13	import	2,622	1,880	0	import[1463] imported[191] importing[968]
14	waste	2,574	3,064	0	waste[2432] wastes[106] wasting[36]
15	disadvantage	2,409	1,648	0	disadvantage[1357] disadvantages[1052]
16	debt	2,363	2,338	0	debt[1825] debts[538]
17	abortion	2,167	2,794	0	abortion[2039] abortions[128]
18	material	1,914	830	0	material[1082] materials[832]
19	parent	1,765	964	0	parent[920] parents[845]
20	job	1,729	1,368	0	job[1317] jobs[412]

In order to bring out the non-topic specific words, a lemmatised version of the key word list was constructed, using Wordsmith 7’s Keywords function and the lemmaslist5 list of twenty thousand lemmas from the BNC taken from the Wordsmith Tools website at http://lexically.net/wordsmith/support/lemma_lists.html. Lemmatisation involves grouping words with different word forms but the same word class, for example *student* and *students* have different forms due to the spelling, but the same word class as they are both nouns. Although many of the words on the list are the same as on the non-lemmatised version, some non-topic words are apparent in the table above, for example *problem(s), accord(ing), advantage(s)* and *disadvantage(s)*. Rather than being an artefact of the topics of the essays, this demonstrates the type of essay, such as problem-solution essay. Items from further down both keyword lists revealed little of interest about the corpora.

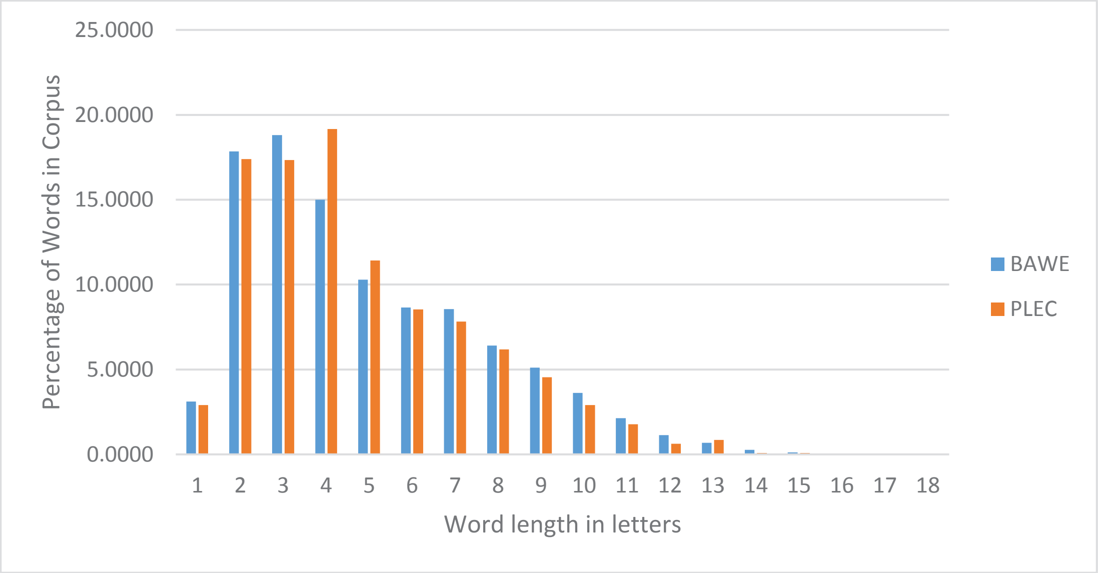
Given the limitations described above, no recommendation for teaching can be made on the basis of the keyword analysis. However, if teachers or students are creating their own corpora, consistency in corpus composition in terms of essay prompt and type of essay is useful for the generation of word lists.

4.4.2 Word Length

Readability measures count words of certain lengths, for example, the measure known as ‘SMOG’ (McLaughlin, 1969) counts words of three syllables or more. Thus if readability is related to proficiency, word length may also be a factor.

The percentages of words of different lengths were compared, as can be seen in Figure 4.4.2 below. The level one English essay writers in BAWE-EON used more one-to-three letters words, and more words with seven or more letters. A possible reason for this is language differences between English and Chinese. In English, some short words such as the articles *a*, *an* and *the*, some short determiners such as *my*, and some short prepositions such as *at* seem to be more frequent than in Chinese, in which they seem to have a less prominent role in sentences. A reason for this is that Chinese does not have articles (Li & Luk, 2017, p. x) or phrasal verbs (Liao & Fukuya, 2004, p. 200). In BAWE-EON the percentages of articles, determiners and prepositions used were 9.99, 14.26 and 16.09 respectively, while in PLEC they were 8.06, 11.94, and 14.19, all less than in BAWE-EON.

Figure 4.4.2: Comparison of Word Length Percentages in PLEC and BAWE-EON



4.4.3 Type-token Ratios

Another text feature that may be related to proficiency is type-token ratio, which is a measure of the lexical diversity or richness of texts (Szudarski 2018; Staples and Reppen, 2016). Cheng (2012, p. 218) defines *type* as each distinct word in the corpus (not including repeats), and *token* as each word in the corpus irrespective of whether it is repeated. A type/token ratio is thus 'the proportion of distinct words and total number of words in a corpus'. Higher ratios indicate that students are not repeating the same word as often as those with lower ratios, thus indicating a wider vocabulary.

Studies such as Staples and Reppen (2016) have found that L1 English speakers use a greater variety of vocabulary than Chinese students, and that L2 writers rely more on repetition of vocabulary. Gui and Yang (2001) investigated the Chinese Learner English Corpus (CLEC), and found that 'The TTR (Type/Token Ratio) of CLEC is much smaller than those of native speakers' corpora, showing that the Chinese learners have a limited vocabulary… The vocabulary of College English students appears to be smaller due to the constraint of topics, because their written works were chosen from examination papers.' Their statistics show that the type-token ration for CLEC was 0.014176, as compared to other corpora with ratios such as 0.049699 for the Brown corpus of American English and 0.029327 for the Lancaster-Oslo/Bergen Corpus (LOB) corpus of British English. However, the composition of CLEC includes texts from middle school students, so analysing PLEC and comparing it to BAWE, neither of which featured in Gui and Yang's investigation, may shed further light on this area.

The essays in BAWE-EON contain 27,303 word types and 674,227 tokens, giving a ratio of 0.040495, between the figures for the Brown and LOB corpora. However, the PLEC EAP sub-corpus contains 12,108 word types and 656,987 tokens, giving a ratio of 0.018429, closer to the CLEC figure.

It is still unclear however, which of Gui and Yang’s reasons: limited vocabulary, task constraints, or both, causes the smaller ratio. To test this, the type-token ratios in the PLEC corpus were investigated further. The corpus contains files grouped according to grades in two assessments: firstly, the students grade on the undergraduate essay assignment on their EAP subject, and secondly on their grades in the school leaving English exam at the time, known as the Use of English ‘A’ level (UEA). The tasks for the students were limited in the same way by the task constraints, but students' vocabulary would not be so limited as it would have related to their language ability. Thus the task constraint factor is removed or greatly reduced because assessment tasks are designed to have similar levels of difficulty for parity reasons.

Table 4.4.3 below shows the type-token ratios of the two assessments and the grades that the students received. The ratio is calculated in two ways, the normal type-token ratio, and the logarithmic type-token ratio, which decreases the effect of different text sizes. Although there are a number of statistical techniques for this, the standardised type-token ratio (STTT) as calculated by Wordsmith Tools 7 using a basis of 1,000 words and logarithmic type-token ratio (log TTR) were selected as Cheng (2012, p. 63) recommends STTR, and log TTR as it is used by Gui and Yang, thus aiding comparability to their findings. Standardised TTR is measured for groups of words; e.g. every 1,000 words, therefore avoiding different text lengths influencing the TTR.

A consistent downward trend can be seen from the table in the UEA type-token ratio of grades A to E, and in the EAP grades from A to C. It thus appears likely that Gui and Yang may be correct, in that students' vocabulary level, rather than task constraints, is a more important factor. However, too much should not be read into these results because vocabulary ability is only one constituent criterion of the students' grades, and other criteria such as grammar may have a greater bearing on the overall grade.

Table 4.4.3: Comparison of type-token ratios across grades in PLEC

Assessment	PLEC-UE					PLEC-EAP				
Grade	Types	Tokens	TTR	STTR	Log TTR	Types	Tokens	TTR	STTR	Log TTR
A	370	997	0.37111	-	0.856440	1848	14193	0.13021	35.78	0.786764
B	1164	4917	0.23673	39.25	0.830499	7127	220912	0.03226	34.81	0.720948
C	3158	34556	0.09139	36.18	0.771046	9043	347846	0.02600	33.89	0.713957
D	4568	92467	0.04940	34.36	0.736958	4410	72774	0.06060	34.79	0.749580
E	6139	157526	0.03897	33.74	0.728851	Grade letter not used				
F	3457	55702	0.06206	33.14	0.745638	417	1262	0.33043	32.70	0.844916
Unclassified	1136	6856	0.16569	32.20	0.796486	Grade not used				

It is difficult, therefore, to draw any conclusion, other than that type-token ratio does seem to scale with proficiency to a limited extent for Chinese learners, and that further research is needed.

4.4.4 Norm-specific Vocabulary

The existence of norms specific to academic fields, and the need to analyse corpora in these terms is highlighted by Nesi and Gardner (2006, p. 114). Breeze (2011) describes a specific example, that ‘reasonable/ly, appropriate/ly, correct/ly’ and ‘proper/ly’ appear to convey attributes that have particular importance in the legal profession.

A sub-sub corpus of essays by level one English native speakers that are tagged as in the field of law was extracted from the BAWE-EON corpus, containing 14 texts comprised of 25,957 words. A total of 21 examples of ‘reasonable/ly, appropriate/ly, correct/ly’ and ‘proper/ly’ were found, which is 0.081 percent. The same search was done on the sub-corpus of all level one essays by native English speakers, finding 300 occurrences in 640,013 words, or 0.046 percent. Therefore Breeze’s finding seems to be correct in the BAWE-EON corpus essays. It was not possible to test this in the PLEC corpus because there is no law department involved.

4.4.5 Academic Word List

The Academic Word List (AWL) (Coxhead, 2000a) is a list of words that are used with higher frequency in academic contexts than in general contexts, and which therefore excludes high-frequency words on West's (1953) General Service List. It is arguably the most widely used EAP word list nowadays taking corpus frequency into account' (Ackermann & Chen, 2013, p. 235).

The AWL is designed to be taught to students who are learning academic writing. Selection of words for the AWL is based on the principles that 'teachers should teach the most useful vocabulary no matter what subject area the students will study in future' (Coxhead, 2000a, p. 73). The AWL is deliberately not discipline-specific. Another design principle of the list is that the most frequent items should be taught first, so the list is divided into frequency-based sub-lists. The derived form of the words is included, for example *accommodate, accommodated, accommodates, accommodating,* and *accommodation* (Sublists of the Academic Word List, 2010). This word families approach has been criticised by Gardner and Davis (2014) for including items with different parts of speech and different meanings, such as *react* and *reactionary*, in the same family.

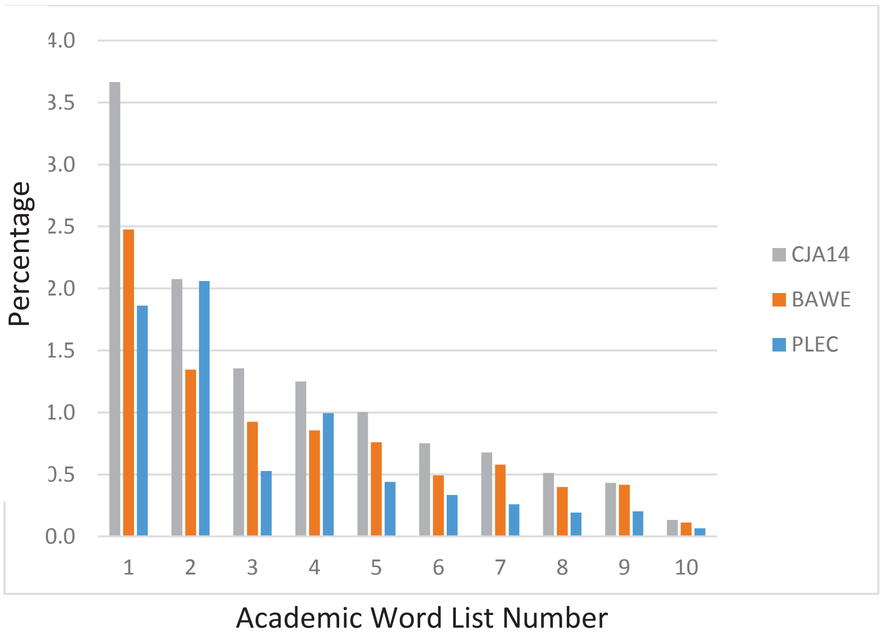
The PLEC, BAWE-EON and CJA14 corpora were searched for AWL words, and from Table 4.4.5.1 and the bar chart in Figure 4.4.5 below it can be seen that the PLEC percentages start at higher levels for the lists of more common words, but in the lists of rarer words the percentages are generally lower than BAWE-EON, which is lower again than CJA14.

The high percentage of List 2 words in PLEC was influenced by the word *credit*, which is in List 2, and credit cards was one of the topics of the PLEC essays, resulting in the word being used over 5,000 times compared to an average of 20 times for the other words in that list.

Table 4.4.5.1: Frequency of the Academic Word List words in PLEC and BAWE-EON

	PLEC		BAWE-EON		CJA14	
List	Frequency	%	Frequency	%	Frequency	%
1	12,232	1.8618	15,851	2.4767	218686	3.6660
2	13,529	2.0592	8,606	1.3447	123799	2.0753
3	3,462	0.5270	5,914	0.9240	80833	1.3551
4	6,529	0.9938	5,479	0.8561	74636	1.2512
5	2,883	0.4388	4,857	0.7589	59718	1.0011
6	2,199	0.3347	3,150	0.4922	44899	0.7527
7	1,695	0.2580	3,701	0.5783	40369	0.6767
8	1,256	0.1912	2,550	0.3984	30587	0.5128
9	1,328	0.2021	2,658	0.4153	25837	0.4331
10	434	0.0661	719	0.1123	7891	0.1323
All	45,547	6.9327	53,485	8.3569	707255	11.8563

Figure 4.4.5: Comparison of the percentages of AWL words in PLEC, BAWE-EON & CJA14



For List 4, two of the most frequent words were *professionals* and *professional*, with frequencies of 1,961 and 467 respectively. The other two words in the top 4 for this list were *job* and *jobs*, at 906 and 412 occurrences. This is against a mean frequency of 22 for words on List 4. The high frequency occurred because one of the essay topics was on the effect of importing non-Hong Kong professionals to work at jobs in Hong Kong. These statistics demonstrate a drawback of using a word list to deduce students' vocabulary size if the corpus texts do not have sufficient range of topics.

This and the overall figures in the last row of the table seem to indicate that PLEC students are using fewer academic words, especially from the higher-numbered lists of rarer words in list five and higher. Therefore it can be concluded that teaching of the AWL should be more emphasised, and methods of implementing this are examined in the Discussion chapter. However, it should be noted that Hyland and Tse (2007) found that some of the items from the AWL vary enormously in different genres across disciplines in terms of range, frequency, collocation and meaning, and therefore for single-discipline classes a more discipline-specific academic word list might be more appropriate.

To investigate Hyland and Tse's finding, and to check that AWL words are not just randomly distributed across corpora, the discipline-specific figures for BAWE-D and CJA14-D were compared. The assumption was that if the words are discipline-specific, their use in the same discipline of different corpora should show a positive relationship. For each of the broad disciplines the normalised frequency of the AWL words in lists 1 – 10 were calculated, and then compared between BAWE-D and CJA14-D using Pearson's correlation. The results are in Table 4.4.5.2 below.

As can be seen in Table 4.4.5.2, there is considerable variability between the means for the different disciplines shown in the bottom row, with AWL word use being most similar in arts and humanities, and least similar in life sciences. Regarding the means for each list, shown in the right-most column, there is a general trend that the high correlations are for the words in the early lists, and lower correlations in later lists, with the exception of List 9. The reason for this list being exceptional is that it contains the word *found* over 400 times in both corpora. If this List 9 is removed, the Spearman r is 0.879 with a p of 0.0017, demonstrating a trend. The trend in correlations may be because the words in the later lists are of lower frequency and the number of words in the list is fewer. Overall the mean correlation for all four disciplines was 0.44, and this result is probably affected by the fact that the BAWE writers are first year undergraduates, very different in experience in writing for their discipline from the academics who wrote the papers in the CJA corpus.

Table 4.4.5.2: Comparison of disciplines for AWL words in BAWE-D and CJA14-D

List	Correlation between BAWE-D and CJA14-D				
	Arts & Humanities	Life Sciences	Physical Sciences	Social Sciences	Mean
1	0.64 (0.000)	0.49 (0.000)	0.67 (0.000)	0.67 (0.000)	0.62 (0.000)
2	0.56 (0.000)	0.36 (0.000)	0.37 (0.000)	0.63 (0.000)	0.48 (0.000)
3	0.62 (0.000)	0.38 (0.000)	0.43 (0.000)	0.54 (0.000)	0.49 (0.000)
4	0.51 (0.000)	0.09 (0.000)	0.20 (0.001)	0.62 (0.131)	0.36 (0.033)
5	0.56 (0.000)	0.14 (0.000)	0.65 (0.000)	0.42 (0.008)	0.44 (0.002)
6	0.49 (0.000)	0.07 (0.000)	0.54 (0.000)	0.47 (0.177)	0.40 (0.044)
7	0.38 (0.000)	0.28 (0.000)	0.25 (0.000)	0.50 (0.000)	0.35 (0.000)
8	0.41 (0.000)	0.27 (0.000)	0.28 (0.000)	0.47 (0.000)	0.36 (0.000)
9	0.72 (0.000)	0.72 (0.000)	0.66 (0.000)	0.63 (0.000)	0.68 (0.000)
10	0.43 (0.000)	0.24 (0.096)	0.23 (0.022)	0.17 (0.018)	0.27 (0.034)
Mean	0.53 (0.000)	0.30 (0.009)	0.43 (0.002)	0.51 (0.033)	0.44 (0.011)

Note: The left number is the Pearson correlation, in brackets is the significance p level

To remove this factor, the disciplines in the CJA14 corpus were compared to the rest of the same corpus, so for example, the arts and humanities discipline was correlated with the sum of the data for life, physical, and social sciences, glossed as ‘vs. others’ in Table 4.4.5.3. It was expected that these correlations would be low if the normalised frequencies of AWL words in the corpus were discipline-specific.

However, as can be seen in the table, the frequencies of the arts and humanities AWL words correlated closely with those in the other disciplines. The least correlation is seen in the life sciences discipline. This tends to add weight to the 'common core' hypothesis (Bloor and Bloor, 1986, p. 5; Flowerdew, 2012, p. 210; Coxhead, 2000b) that learning of general English takes first place chronologically in language learning. However, the table only gives information about the frequency of use of the AWL words, and does not address meaning or collocation.

Table 4.4.5.3: Comparison of AWL words in the disciplines of the CJA14-D corpus

AWL List	Correlations				
	Arts and humanities vs. others	Life sciences vs. others	Physical sciences vs. others	Social sciences vs. others	Mean
1	0.99	0.90	0.88	0.88	0.91
2	0.97	0.77	0.88	0.80	0.86
3	0.99	0.84	0.89	0.79	0.88
4	0.99	0.77	0.88	0.77	0.85
5	0.99	0.83	0.92	0.67	0.85
6	0.97	0.57	0.87	0.74	0.79
7	0.97	0.63	0.87	0.78	0.81
8	0.98	0.75	0.89	0.77	0.85
9	0.99	0.91	0.95	0.92	0.94
10	0.99	0.73	0.71	0.84	0.82
Mean	0.98	0.77	0.87	0.80	0.86

Note: All correlations in this table were significant to the p = 0.000 level

Regarding range, there are words that only appear in one discipline, for example *unassessed* from List 1 of the AWL only appears in Life Sciences, and *conceptualisation, conceptualise, conceptualised* and *conceptualising* from the same list do not appear at all in Life Science, but do in the other disciplines. Table 4.4.5.4 below gives details of the range with which words are missing in from at least one but not all disciplines in each AWL list.

As can be seen in Table 4.4.5.4, there is considerable variation in the percentage of words from the AWL lists that do not occur in a discipline, with a mean of 13% of AWL words from each list for the Life Sciences discipline not appearing in the corpus. Therefore, for single-discipline classes it is still the case that a more discipline-specific academic word list might be more appropriate.

Table 4.4.5.4: Range of AWL words in CJA14-D

AWL List	Percentage of Missing AWL Words in CJA14-D Disciplines				
	Arts and humanities	Life sciences	Physical sciences	Social sciences	Mean
1	1.32	21.41	9.71	5.08	9.38
2	1.99	15.45	9.49	3.75	7.67
3	0.44	11.70	6.84	3.09	5.52
4	1.32	9.49	5.74	2.43	4.75
5	2.21	15.67	9.93	4.42	8.06
6	0.88	16.56	7.06	6.18	7.67
7	0.22	13.02	4.19	2.21	4.91
8	1.32	14.13	8.83	4.19	7.12
9	0.00	11.04	5.96	4.19	5.30
10	0.66	5.30	1.55	1.32	2.21
Mean	1.04	13.38	6.93	3.69	6.26

Another reason for the lack of some AWL words in the CJA14 corpus might be that the AWL's coverage is unrepresentative. However, Coxhead took pains to ensure that the AWL was representative, as words were extracted from a balanced corpus selected from academic articles, university text books, two laboratory manuals, and parts of corpora, including the Wellington Corpus of Written English, the Brown corpus, the LOB corpus, and the MicroConcord corpus, equally distributed across arts, commerce, law and science (Coxhead, 2000, p. 220). The selection criteria of words for the list were specialised occurrence, in that they had to be outside the General Service List (GSL) (West, 1953); their range had to be ten or more times in each section of the corpus and in 15 or more of the 28 subject areas; and words in a 'word family' comprising the headword and its derived forms had to occur a minimum of 100 times in the corpus. Coxhead double-checked the list against both a second corpus that she built, and the University Word List (Xue and Nation, 1984).

However, the text selection method for these corpora was described as 'opportunistic' by Hyland and Tse (2007, p. 239) and despite the approximately equal number of words in Coxhead's disciplinary sub-corpora, they criticise it based on the uneven number of texts that contain these words. Regarding coverage, they point out that it is not evenly distributed, and combining the AWL and GSL misses covering 22% of the words in their corpus. However, this is not surprising, because earlier in their paper they describe English vocabulary as having three parts: high-frequency words such as those in the GSL, academic vocabulary such as in the AWL, and also technical vocabulary, which differs by subject area and covers up to 5% of the words in texts. Therefore, with multiple texts from different disciplines in their corpus, one would expect considerable overlap between the GSL and AWL words in each essay, but also for there to be a distinctive technical vocabulary in each essay, based on discipline and topic, which would add up from each essay to form a proportion of the missing words in the corpus.

To shed light on this issue, a separate academic word list was used to test the CJA14 corpus. This word list is the Academic Keyword List (AKL) by Paquot (2010). The details of the methods by which she selected the items on the list are given in Paquot (2010, pp. 44-55).[6] The CJA14-D discipline-specific sub-corpora were then searched for these words, and 3,090,307 tokens of the AKL-F list words were found in the CJA14 corpus, against a total number of tokens in the corpus of 5,965,229, or a coverage of 51.8%, which is far less than the AWL coverage of 22% missing cited

[6] In summary, Paquot used keyness to extract key words by comparing a corpus of literature from well-known corpora with a corpus of student writing. The keywords were then filtered for having sufficient range and evenness of distribution, and then more were selected if they were related to those already in the list. She helpfully provides the headwords list online, and to give comparability to the AWL word families were needed, so for this analysis the words were taken, the noun list had plural forms added, and the verb list had the third-person, continuous past and perfect forms added, and named the AKL-F list (AKL with word families).

by Hyland and Tse (2007). To ascertain whether any of the words in the AKL-F were discipline-specific, in that they occurred with a frequency of zero in one discipline, but were present in the other disciplines, a comparison was done and the results are in the Table 4.4.5.5 below.

A comparison of the table above with the similar Table 4.4.5.4 above on the AWL shows that there is much less disciplinary variation with the AKL-F, with a mean of 0.47% words missing in at least one discipline, than the AWL, which had a mean of 6.26%. There are a number of possible reasons for the difference. The first may be the rarity of some of the words that Coxhead included in the AWL, for example forms with both British and American spellings such as *uncontextualized* and *uncontextualised* (Coxhead, 2000b). Another possible reason for the difference is the number of words in the list: AWL has 3,112 words in total, but the AKL-F has 1,961 words, so there are less opportunities for rare words to be missing in the AKL-F. In addition, the AKL-F differs from the AWL due to the inclusion of high-frequency words such as *the, in, aim, because, explain* and *result*, which cover 24% of the tokens in the CJA14 corpus. Despite these factors, the figures for the number of words in the AKL-F list that do not appear in the CJA14-D corpus are much lower than for AWL, showing that the AKL-F words are more common in academic texts.

Table 4.4.5.5: Missing Words from the Academic Keywords list in CJA14-D Disciplines

AKL-F Components	Percentage of Missing AKL Words in CJA14-D Disciplines				
	Arts and humanities	Life sciences	Physical sciences	Social sciences	Mean
Nouns	0.00	2.64	2.03	0.81	1.37
Verbs	0.11	4.31	0.75	0.32	1.37
Adjectives	0.00	0.00	0.41	0.00	0.10
Adverbs	0.00	0.00	0.00	0.00	0.00
Others	0.00	0.00	0.00	0.00	0.00
Phrases	0.00	0.00	0.00	0.00	0.00
Mean	0.02	1.16	0.53	0.19	0.47

There is also little difference between the disciplines, with a maximum of less than 4.31% of the verbs being missing in the life sciences discipline but present in others, thus lending support for the ‘common core’ hypothesis. This may be because the AKL words are more frequent in English, as the GSL words are not excluded.

Other academic vocabulary word lists are available.[7] In future, as corpus sizes increase and statistical methods become more refined, it is probable that more such lists will be produced, each claiming to be an improvement.

Another area of concern regarding vocabulary lists is the difference between vocabulary needed for receptive purposes such as reading and listening, and vocabulary needed for productive purposes such as essay writing and presentations. This issue was investigated by Malmström, Pecorari, and Shaw (2018), who analysed the whole BAWE corpus using the AVL, and found that students’ productive vocabulary was six times smaller than the whole AVL, the pedagogical implication being that the productive vocabulary should have a central position in teaching because learning it takes longer than learning vocabulary for receptive purposes only, due to the need to integrate the vocabulary appropriately with the surrounding text and context.

Recommendations for teaching include that because the AKL list contains higher-frequency words and little disciplinary variation, it is suitable for lower level and multi-discipline classes. The AWL is suitable for more advanced classes, especially

[7] An example of another vocabulary list is the Academic Vocabulary List (AVL) by Gardner and Davis (2014), which was compiled by an analysis of the 120-million word academic sub-corpus of the Corpus of Contemporary American (COCA) corpus, giving 3,014 lemmas in the list. However, the AVL was found by Durrant (2016) to contain significant variation across disciplines, with only a smaller core of around 400 items being a generic productive academic vocabulary.

for arts and humanities and social sciences classes, but perhaps less so for classes in the life sciences discipline. Another possibility is for the teacher or subject leader to produce their own list, because the contents of such lists depend firstly on the corpora that they are taken from, which is in turn dependent on the availability of suitable texts and design choices taken when compiling it, such as genre and author background, and secondly the design choices made when selecting words for the list, for example, regarding range and evenness of distribution (Paquot, 2010, p. 45).

There are also practical concerns regarding the use of the lexical items in class, such as the inclusion of words that cause common errors, such as *lack* for Cantonese L1 learners, and words that are wrongly used in a stage of the learners' interlanguage, such as *hardly* being used as the opposite of *easily* by Cantonese L1 learners. Therefore a class-specific academic word list could be compiled by comparing a suitable pedagogical corpus with a general corpus, extracting the keywords, creating word families around them, and then filtering and prioritising them for range, distribution and teachability.

4.4.6 Academic Formulas List

Although academic word lists are well-known in the literature (Flowerdew, 2012, p. 192; Cobb and Horst, 2015, p. 190), they are limited in that they tend to contain single words only, rather than incorporate groups of words that commonly appear together, and which students could use as chunks, which would help with fluency and native-like word selection (Simpson-Vlach and Ellis, 2010, p. 488). The importance of phrases for language learning was highlighted by Sinclair (1987) who stated that writers co-select the words that they use through semantic prosody, which is the overall attitudinal and pragmatic meaning of the lexical item, through collocation with other words that are usually found with a word, and through colligation, which concerns the co-occurrence of grammatical choices (Cheng, Greaves, Sinclair and Warren, 2009, p. 239).

Two opposing principles that may explain word choice are explained by Sinclair (1991, pp. 109-10), which are the ‘open-choice principle’ and the ‘idiom principle’. In the open-choice view, a word can be followed by a large number of possible choices of word, restricted only by grammar. In the idiom principle view, a word can be followed by ‘a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments’ (p. 110). These phrases have a number of features, including indeterminate extent due to varying degrees of collocation with other words, lexical variation such as word order and form of pronoun, colligation with tenses, and positive and negative connotations.

The relevance of this to language teaching and language use in academic writing is that readers are prepared for and sensitive to such idioms, according to Hoey’s (2005) theory of lexical priming. Therefore, in order for an academic writer to demonstrate membership of the academic community, it is not enough to use single academic words, but those words should be used in phrases that exhibit the features of the idiom principle.

Examples of phrases following the idiom principle can be found in the Academic Formulas List (AFL), which is a list of common formulaic sequences of three to five words in length in academic English, developed by Simpson-Vlach and Ellis (2010) and designed to be comparable to the AWL. The source of the sequences is Hyland's (2004) 1.5-million word research article corpus of 1,426 texts from five genres in eight disciplines, and also from selected British National Corpus files sampled across academic disciplines. The criteria for the sequences are that they are frequent recurrent patterns in corpora of written and spoken language, which occur significantly more often in academic than in non-academic discourse, and inhabit a wide range of academic genres. The purpose of the list is EAP instruction (Simpson-Vlach and Ellis, 2010, p. 487).

In this research, the sub-list on written academic language is used. This includes sequences such as *on the other hand, due to the fact that, it should be noted, it is not possible to,* and *a wide range of.* Due to the selection criteria of three to five word long sequences, and following Sinclair's feature of indeterminate extent due to varying degrees of collocation with other words, some of the sequences are very similar, for example, *on the other, on the other hand,* and *on the other hand the* are counted as different sequences, which means that the sequence *On the other hand* in the PLEC A+ grade essay matched the list twice, and instances of '*on the other hand, the'* in the PLEC A grade essays matched three times each. For consistency and comparability, this method of counting was applied to both corpora.

The PLEC EAP corpus was compared with BAWE-EON, and as Table 4.4.6 below shows, when the occurrences are normalised by the word count of the sub-corpora, the PLEC student writers underused the formulaic sequences.

Comparing the two corpora, the log likelihood of the difference between them is 5658.02, significant at p < 0.0001, the percentage difference is 303% and the effect size using log ratio is -2.01.

Table 4.4.6: Comparison of Academic Formula Sequence Use

Corpus	Grade	Frequency	Word count	Percent
BAWE-EON	All	11473	640013	1.79
PLEC EAP	A+	3	471	0.64
	A	99	13776	0.72
	B+	322	68041	0.47
	B	688	152981	0.45
	C+	943	214136	0.44
	C	560	133881	0.42
	D+	222	53410	0.42
	D	81	19386	0.42
	F	6	1257	0.48
	Total	2924	657339	0.44

Comparing PLEC to the frequency figures given in the AFL, the PLEC writers overused the following sequences by a frequency of twice or more: *less likely to, they do not, the other hand, it is difficult, on the other hand, on the other, should not be, it is obvious that, the most important, it is necessary, a large number, a large number of*, and *that it is not*.

The reason for the lower percentages in the table above could be explained by two findings. The first is the fact that there were 48 sequences in the AFL that were not used at all in PLEC. These included *take into account the, as can be seen, in this paper we, in the next section, the United Kingdom, on the basis of the, the same way* *as, with respect to the,* and *in the present study*. These expressions were selected as examples because they are sequences with a Formula Teaching Worth (FTW) of over one. FTW is a measure of usefulness created by Simpson-Vlach and Ellis, who describe it as an "empirically derived psychologically valid measure of utility" (p. 488). Its components are statistical and human judgement. The statistical components include frequency and Mutual Information (MI) scores that assess the degree to which the words in a phrase occur together more frequently than would be expected by chance. The human judgements were by language instructors and testers, and considered whether the phrase constituted a formulaic expression, has a cohesive meaning or function, and was worth teaching (Simpson-Vlach and Ellis 2010, pp. 493-6). At least one, *the United Kingdom*, is probably included because the phrases were taken from the British National Corpus, and would not be expected to be so prominent in Hong Kong students' writing.

The second reason for the lower percentages in Table 4.4.6 is that the PLEC students underused almost all of the sequences compared to the writers of BAWE-EON. Using Wordsmith 7's concord programme, a concordance of the sequences was taken for both corpora, and the resulting sentences were then analysed using Xu's (2009) log likelihood calculator. Of the 200 sequences, the log-likelihood statistics show that all but 15 were significantly underused by the PLEC writers compared to BAWE-EON. These 15 sequences were: *should also be, on the other hand the, the other hand the, to determine whether, a large number of, are as follows, should not be, a large number, it is obvious that, carried out in, depend on the, if they are, depends on the, to carry out, be used as a,* and *whether or not the.* It should be noted that some of these are very similar, such as *on the other hand the, the other hand the* and *depend on the, depends on the*. There is also one unhedged sequence, *it is* *obvious that*. The sequences and statistics can be seen in Appendix Two.

Based on these findings, input to students on these frequent, wide-ranging, teachable and teaching-worthy sequences should be considered, and suggestions regarding how this could be done are detailed in Chapter Six.

4.4.7 Repertoire of Recurrent Word Combinations

A general pattern found in phraseological research, such as for the AFL, that 'non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers' even when the non-native speakers are of an advanced level, was found by Adel and Erman (2012, p. 90). In addition to over-use of AFL phrases, they found that native speakers tended to use more unattended 'this' constructions, existential 'there' constructions, hedges, and passive constructions.

The corpora were search for examples of these constructions. Searching for this constructions using *this is/was/has been/had been/will be* found 426 in the BAWE-EON, (0.0666%), compared to 115 for the PLEC (0.0175%). To compare to advanced level non-native speakers a sub-corpus of A and B grade scripts was extracted from the PLEC corpus, containing 235,467 words. In this 38 instances (0.0161%) were found, seeming to support the authors' conclusion. However, in my experience Hong Kong students are often unaware that they should use 'this' instead of 'it' when referring to a situation. Therefore the search was modified to look for *this/it* followed by *is/was/has been/had been/will be*, and found 2,160 in the BAWE-EON, (0.3375 %), compared to 1,740 for the PLEC (0.2648 %), and 609 instances (0.2586%) were found in the sub-corpus of A and B grade scripts, seeming to support Adel and Erman's conclusion about unattended 'this' constructions. Students can be taught how to use these constructions, for example for giving reasons using *This is because.*

Searching for the existential 'there' tags in the tagged versions of the corpora found 1,665 in the BAWE-EON, (0.2369%), less than the 2,265 for the PLEC (0.3265%), and 621 instances (0.2637%) were found in the sub-corpus of A and B grade scripts, seeming to contradict the authors' conclusion. This may be a special case because the Cantonese verb for *to have* (有jau5) can be translated as an existential 'there' construction at the start of a sentence, and the simplicity of this translation may incline students towards using it more frequently. There is support for this in Larsen-Freeman and Long (1991), in which they discuss how interlanguages are influenced by the learners first language (pp. 96-107), and use an example from Schachter and Rutherford's (1979) research, in which a Chinese subject overproduced existential sentences with a dummy subject, which they hypothesized 'was due to the learners' having seized a particular English syntactic pattern to serve a discourse function that their L1s, being topic-comment, require' (p. 97).

The use of existential *There* by Chinese students was studied in detail by Lin (2003, p. 289), using almost 4 million words of student essay assignments by students in Hong Kong as the specialised corpus and LOCNESS as the reference corpus. She concluded that her 'study has provided empirical support for Rutherford's theory regarding the reasons for the production of CIL (Chinese inter-language) *there be* sentences, demonstrating that CIL *there be* construction is a product of typological transfer and the Chinese word *you* is the vehicle of such typological transfer' (p. 289) {*you* being the transliteration of the Putonghua pronunciation of the verb *to have*}.

However, a point related to teaching is that a concordance search for *There is/are/was/were/have* of the PLEC-EAP corpus brought to light a large number of errors relating to subject verb agreement, examples of which can be seen in the figure below. This construction seems to be popular among students, with 487 concordance hits in the PLEC-EAP corpus compared to 356 in BAWE-EON (a log-likelihood of 17, significant to $p < 0.0001$, with a log ratio of 0.41).

Figure 4.4.7: Errors in the use of sentence initial 'There' in the PLEC-EAP Corpus

There are a few similarity between them.
There are a number of benefit,
There are a lot of advantage of
There are a large elderly population in
There are a lot of information in
There are a numbers of pros
There are a lot of incorrect informations
There is 14% of respondents who are mostly
There is not enough professionals in Hong
There is no any prosperous cities
There is much difficulties for the poor
There were a lot of city development
There is some people support the proposal
There were 53 bartender worked in a

This popularity may be due to the simplicity of its translation, however, learners should be advised of the necessity of making the verb agree with the following noun; e.g. 'There are a lot of information' should be 'There is a lot of information', and that the verb becomes the main verb in the sentence, so a relative pronoun may be necessary; e.g. 'There were 53 bartender worked in a…' should be 'There were 53 bartenders who worked in a…'.

Adel and Erman's (2012, p. 90) found that native speakers tended to use more hedges, which are words that express caution about claims by using parts of speech such as modals, adjectives and adverbs of possibility. This was investigated using the following regular expression to search the corpora: \b(this|it)\s+(finding|result|figure)?\s*(may|might|could|is)\s*(probably|possibly|perhaps|maybe|apparently|seemingly|presumably|conceivably|generally|largely|primarily|for\s+the\s+most\s+part|predominantly|mainly|usually|to\s+a\s+great\s+extent|(a|an|one)\s+(possible|probable|unlikely\sto))?(\s*be)\s*?(a\s+(result|consequence)\s+of|because|due)\b , which is unlikely to catch all instances of hedging, but did catch some. It found 1 in BAWE-EON, (0.0002%), 1 for the PLEC (0.0002%), and 1 instance (0.0002%) was found in the sub-corpus of A and B grade scripts.

Therefore a simpler expression was tested: \b((suggests?|indicates?|it (may be|might be|could be|is) (possible|probable|likely))\s+that|likely|largely|rarely)\b .
It detected 592 instances (0.0901%) in PLEC and 668 (0.1044%) in BAWE-EON, slightly more in the latter. This seems to support the authors' conclusion. However, as the low number of examples found for the first regular expression and the close results of the second show, automatic detection of hedging is difficult, because, as Flowerdew (2012) states 'the lexico-grammar realising hedging is extremely subtle, encompassing many different types of lexical verbs and modals' (p. 15) as well as adjectives and adverbs, and therefore no strong conclusion can be drawn.

The use of Adel and Erman's (2012, p. 90) final category, passives, was investigated with the regular expression (is|are|was|were|has been|havebeen)\s+(\w+y\s+)* \w+ed\s+by which looks for a form of *is*, optionally a word ending in 'y' such as 'greatly' and 'heavily', a word ending in 'ed', finishing with 'by'. This is limited because it will not catch irregular verbs or agentless passives. The results were manually filtered to eliminate non-passives. It found 668 in the BAWE-EON, (0.1044%), compared to 355 for the PLEC (0.0540%), and 132 instances (0.0561%) were found in the PLEC sub-corpus of A and B grade scripts, these comparable figures therefore again seeming to support the authors' finding that native speakers use more passive constructions. Passives, both by-passives and agentless passives, are also investigated below in Section 5.3.4 on Dimensions of Linguistic Variation in Register and Genre, and for Dimension Five it was found that the use of both types of passive was higher in BAWE-EON than PLEC-EAP, and also that use scaled significantly with PLEC-EAP grade, with higher-graded texts using a greater proportion of passives.

Therefore Adel and Erman’s (2012, p. 90) finding that native speakers tended to use more unattended ‘this’ constructions, existential ‘there’ constructions, hedges, and passive constructions seems to be partly borne out by the evidence from the PLEC corpus, with the exception of existential ‘there’ constructions, which seems to be a special case due to language transfer.

4.4.8 N-grams

Although both n-grams, word combinations and lexical chunks are multi-word units, the difference is that n-grams are discovered computationally, while lexical chunks are ‘psychologically whole’ (Leedham, 2014, p. 12). For example, the n-gram *allows the reader to* from the research cited below is not psychologically whole because it should be followed by an infinitive verb.

The top n-grams from English studies essays in BAWE and general academic prose taken from academic prose and written fiction in BNC-Baby, which is a sub-set of the British National Corpus, were contrasted by Ebeling (2011). He defines n-grams as groups of words which follow each other in a text (2011, p. 55), and is mainly concerned with three- and four-word combinations. Examples of these n-grams include *allows the reader to, as can be seen, at the beginning of, by the use of, can be seen in, could be argued that, it could be argued, the beginning of the, the importance of the, through the use of, to the fact that,* and *way in which the* (Ebeling, 2011, p. 60). Other n-grams seem to be text- or topic-specific, such as ‘heart of darkness’ and ‘the good soldier’, and are therefore not included in this analysis. These n-grams comprise lexical items with grammatical, cohesive and stylistic aspects such as use of the passive voice, modal verbs, connectives and hedging; aspects which Evans and Morrison (2012) found are ‘generally not included in assessment criteria and thus received less attention in the planning and production of assignments than information and ideas’ (p. 41).

These n-grams were searched for in the corpora. In the BAWE-EON, 225 were found (0.0352 %), a much higher figure as compared to 12 for the PLEC (0.0018%). A reason for the difference may lie in grammar. Ebeling analyses the structure, stating, ‘The most common pattern among the 20 quadrigrams found in academic prose is PREP + (DET) + N + PREP + (DET), e.g. “on the basis of.” ’ As Chinese is very different from English in terms of prepositions and determiners, for example because there is no distinction between definite and indefinite articles, students may not naturally use them, or may avoid them for fear of making mistakes.

In order to test this, the pattern was searched for in special versions of the corpora consisting of tags only. A total of 9,723 instances were found in the BAWE essays by native speakers (1.5192%), compared to 6,053 for the PLEC (0.9213%). These percentages seem to indicate support for Ebeling's finding.

Therefore it seems that Ebeling’s conclusions seem to apply to a comparison of the essays by English native speakers in BAWE-EON and the students scripts in PLEC.

Computationally-discovered n-grams were also the focus of Chen and Baker’s (2010) comparison of expert, native-speaker learner and Chinese learner academic writing that used BAWE for its learner corpora, although the authors follow Biber in terming them ‘lexical bundles’. They define lexical bundles, n-grams, clusters and recurrent word combinations as all being ‘continuous word sequences retrieved by taking a corpus-driven approach with specified frequency and distribution criteria’, and state that the ‘retrieved recurrent sequences are fixed multi-word units that have customary pragmatic and/or discourse functions’ (p. 30).

Chen and Baker’s corpora parallel the corpora in this thesis, but with a number of differences. Their expert corpus is the academic prose category of the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus, called FLOB-J, containing eighty 2,000-

word excerpts from published academic texts, retrieved from journals or book sections, totalling 164,742 words. For the learner corpora they used essays produced by L1 Chinese students of L2 English, called BAWE-CH, with 146,872 words, and a comparable dataset contributed by peer L1 English students, called BAWE-EN, with 155,781 words.

To find their clusters, they used Wordsmith 4, and then filtered the results to remove proper nouns and context-dependent bundles. This resulted in 80 bundles for BAWE-CH, 103 for BAWE-EN and 108 for FLOB-J. For this thesis these 291 bundles were searched for in PLEC-EAP, BAWE-EON and CJA14 using Wordsmith 7 in order to compare Chen and Baker's results to these corpora.

For each bundle and each corpora a normalised frequency per hundred thousand words was calculated, and then the results were correlated. As would be expected, the strongest correlations were between the bundles in BAWE-CH and PLEC-EAP, at 0.71, BAWE-EN and BAWE-EON at 0.45, and FLOB-J with CJA14, at 0.68. That the correlations were not higher is probably because Chen and Baker's corpora were not restricted to essays.

Chen and Baker also found that the use of formulaic expressions grows with writing proficiency. However, there is controversy in some of the literature on this, such as in Hyland (2008a), who stated that, among Master's and PhD students, there was 'a greater reliance on formulaic expressions by less confident or proficient students in constructing their texts' (p. 60), although Hyland did not filter out proper nouns. Paquot and Granger's (2012) review of the literature on the subject states that 'the overall number of lexical bundles tends to decrease as proficiency in the language

…increases' (p. 9). In this thesis, the mean number of the normalised frequencies of the bundles found by Chen and Baker was measured, and the PLEC-EAP corpus had the lowest figure, with a mean of 1.85 repetitions of each bundle, followed by 2.33 for BAWE-EON and 4.34 for CJA14. Thus Chen and Baker do seem to be correct with these corpora, and therefore, as they say 'after careful selection and editing, the frequency-driven formulaic expressions found in native expert writing can be of great help to learner writers to achieve a more native-like style of academic writing, and should thus be integrated into ESL/EFL curricula' (p. 44). However, because this finding is at odds with some of the literature, the selection and editing process is important, as is checking the frequency of the expressions in the type of writing that the students are expected to do. For example, *on the other hand* is the top most-frequently used n-gram in BAWE-CH and third in FLOB-J, but as discussed in this thesis in the section on connectors, it should not be over-used and it should be used correctly in terms of its contrast function.

An Academic Collocations List (ACL) is described by Ackermann and Chen (2013), and was derived from the 25-million word Pearson International Corpus of Academic English. It contains 2,469 collocates, selected by a combination of corpus-driven and expert judgement techniques. The PLEC-EAP and BAWE-EON corpora were searched for these collocates, and almost 50% were found in neither corpus, 5% were in PLEC-EAP but not BAWE-EON, 35% in BAWE-EON but not PLEC, and 7% were in both, but with only 2% being significantly different. In CJA14, 55% of the collocates were not present, and 20% occurred only once. Given that over 1,200 items were not found in either the PLEC-EAP or BAWE-EON corpus, the large number of items in the list, and that the list is also not ordered for frequency or teachability, there are concerns with the usefulness of this list for teaching.

4.4.9 Readability

In Evans and Morrison's study, graduating students were cited as being dissatisfied with "their unsophisticated writing style, limited repertoire of sentence patterns" (2012, pp. 40-1). One method that can be used to examine the sophistication of writing and the length of sentences is to examine readability statistics, as these are based partly on sentence length.

In this study, readability statistics were analysed using two programs, called Flesch 2.0 (Frink, 2007) and RocketReader Readability (Ronald, 2013), both freely available at sourceforge.net . Two programs were used because readability statistics are based on word and sentence counts, and different software counts these differently. A variation of 0.5 to 3.7% was found between the programs for different measurements. This seems to be partially caused by the various options for counting. In Flesch, these options include counting semi-colons and colons as end of sentences, and counting words found in titles and headings. Only the former was selected. This is probably why sentence counts were 3% lower in Flesch than RocketReader Readability. A test was done with all the Flesch options turned on, which increased the word and syllable counts, probably due to including the headings, but did not increase the sentence count, which remained less than in RocketReader Readability.

There were also differences between the measurements of the two readability programs compared to the figures provided in the BAWE documentation. For example, the documentation's Excel gives a word count of 640,013 words for the level one English essays, but Rocket Readability gives 657,460 and Flesch gives 662,781: differences of 2.7% and 3.5% respectively. This is probably because the programs count words in different ways, and count different stretches of text as words, for example, digits.

However, the findings described below demonstrate that the differences in readability scores between the PLEC and BAWE-EON essays were large enough that these differences were overshadowed.

As mentioned above, a number of sub-corpora were created for the study. For the BAWE-EON corpus these include one of all the essays by native English speakers, and subject-specific sub-corpora of American studies, archaeology and classics essays. The PLEC corpus is already divided into sub-corpora based on the Hong Kong UE exams, with a range of grades, and also on the 22 academic departments that the students were from. A sub-corpora was then created by concatenating all these departmental corpora together into an all-department sub-corpus.

Readability statistics were measured using the various readability programs.[8] Their output is in number of years of education necessary to understand a piece of writing. To compare of the corpora the mean and range of these readability results from different programs were calculated.

There was an appreciable difference between BAWE-EON sub-corpus and the PLEC all-Departments sub-corpus in terms of the results, which are given as equivalent years of education in the American educational system, with the former averaging 15 years of education with a range of 3, and the latter averaging 12 years of education, also with a range of 3 years. The former showed about 25 words per sentence, 5.3

[8] The Flesch program only provides results for Flesch Kincaid Grade Level and Flesch Reading Ease, but RocketReader Readability also provides the Gunning Fog Readability Index, Simple Measure of Gobbledygook (SMOG), and Coleman-Liau Readability Index. Although these measures of readability are all calculated in different ways, their output (except for Flesch Reading Ease) is in number of years of American education necessary to understand a piece of writing.

characters per word, and 1.78 syllables per word. The latter corpus showed far fewer average words per sentence, at 16; and slightly shorter words on average, with 5.13 characters and 1.63 syllables per word. Thus the BAWE-EON students' sentences were about 56% longer. There was also a marked difference in sentence length in the other sub-corpora. The sentence length statistics for the BAWE-EON texts seem to be generalisable, as Hartley (2008, p. 163) found that academic student essays in his research averaged about 25 words in length. If the BAWE-EON essays are to be taken as a model for Hong Kong students as represented by the PLEC students, longer sentences would be recommended.

To ascertain whether the differences in mean sentence length were statistically significant, a z score test was used (Oakes, 1998, p. 9). The mean sentence lengths and standard deviations were calculated by Wordsmith 7's Wordlist statistics tool. In order to compare like with like, it was necessary that the EAP corpus PLEC files, which consisted of groups of essays for each grade letter, be split into individual files using Wordsmith's File Utilities Splitter function. A total of 1,233 files were generated, and to check whether these files consisted of one essay each, the file lengths of all were compared, and then a sample were opened. All were found to be of one essay. Mean sentence length and standard deviations were calculated, and the z scores generated using the "Z-test for two Means, with Known Population Standard Deviations" online calculator at http://mathcracker.com/z-test-for-two-means.php. The mean sentence lengths were 24.68 and 17.92, and the standard deviations were 12.55 and 10.23 for the BAWE and PLEC sub-corpora respectively. The z score was 9.017 and the p value was zero, indicating that the null hypothesis was rejected, meaning that the sentence lengths were significantly different.

The longer sentence lengths in BAWE-EON may be due to the students' level of proficiency. Brown (1973) posited a "Mean Length of Utterance" to measure language development, although he was working on child language acquisition by native speakers. If sentence length is a predictor of higher grades, as the PLEC grades show in that essays awarded an 'F' grade averaged 12 words per sentence, those with a 'C' 16, and those with an 'A' 18 words, more training for students in general proficiency may be beneficial, assuming that sentence length is dependent on proficiency. The L2 Syntactic Complexity Analysis described above, although performing the statistical analysis in a different way, generated similar differences between the mean word lengths of different PLEC grades (see Appendix One). Mean sentence length increased with PLEC grade, as did mean length of T-unit, amount of subordination, and verb phrases per T-unit. In addition, in the 150 BAWE-EON essays, scores were higher than PLEC scores for all of these. In summary, the relationship between mean sentence length, some measures of syntactic complexity, and proficiency seems to be supported.

In addition, if sentence length is also partly a measure of rhetorical sophistication that convinces a reader of the quality of the content, for example by expressing the writer's ideas in a genre-typical manner, it may also be helpful for students to use pre-set academic expressions which can contribute to sentence length, such as those suggested in Morley's (2014) University of Manchester Academic Phrasebank.

4.5 Summary

In this chapter existing teaching materials were initially analysed in order to provide an organisational structure for further analysis. The first main area in this structure was grammatical features of the students' writing. Findings that have implications for course materials for students include those on the use of personal pronouns, and learner strategy training on proof-reading, including with the aid of computers. However, some university lecturers in the UK who were interviewed by Leedham (2014) stated that "as long as they could understand the writing then grammatical errors were 'not usually a problem' " (p. 115), while others were concerned that punctuation and spelling errors reflected students' lack of attention to detail.

Regarding the second main area of the organisation structure, vocabulary features, analysis of word lengths showed differences in the distribution of words of both short and long length. A possible limitation of the students' English was the limited vocabulary revealed by the analysis of standardised type-token ratio, which could also be addressed by course material. Analysis of the use of items from the Academic Word List showed that Hong Kong students tend to use less of the words from the higher-level sub-lists of the AWL, and course materials could give further instruction in this area.

In the area of phraseology, regarding the Academic Formulas List, the PLEC students under-used over 90% of the sequences, and more attention to these sequences could be given in teaching. In addition, analysis of recurrent word combinations indicates that students can be encouraged to use more unattended 'this' constructions and passives, and analysis of n-grams suggests that students could use more of Ebeling's academic phrases.

Readability analysis showed significant differences between the corpora, especially in mean sentence length, which may relate to students' dissatisfaction with the sophistication of their writing and limited repertoire of sentence patterns and could be addressed with course materials.

While this chapter has examined features categorised under language in the comparison of language features in existing materials, the next chapter continues the description and analysis of the findings in the areas of content, organisation and conventions.

Chapter Five: Findings on Content, Organisation and Conventions

This chapter continues to describe and analyse the findings of the research, examining the areas of content, organisation and conventions. The section on content covers phraseological profiles and disciplinary variation. The section on organisation looks at connectors, and the section on conventions describes referencing and citation, rhetorical questions, register, dimensions of linguistic variation, and finally stance and voice.

5.1 Content

The investigation of content by corpus linguistics techniques such as concordancing and frequency counts is limited by the inability of computers to understand the meaning of a text, and judge whether that meaning is correct. However, it is possible to examine some aspects related to content by examining whether the phrases in a text are functional or lexical, and whether the text shows different characteristics when it is written by writers from varying disciplines.

This section therefore investigates the distribution of discipline-specific language by examining the phraseological profiles of the students' writing, and examines disciplinary variation in the BAWE essays and PLEC corpus. The aim of the section is to assess whether there is significant variation between disciplinary writing that would warrant it being included in teaching and learning materials. It should be noted that the content of the two corpora is different, the BAWE essays are on disciplinary topics in arts and humanities, life, physical and social sciences, whereas the PLEC essays were written on topics given by the students' English teachers on topics such as smoking, restaurants, immigration and recycling. This limits the type of comparisons that can be done. Despite this there were some interesting findings.

5.1.2 Phraseological Profiles

Cheng, Greaves, Sinclair and Warren (2008) state that it is important to examine the phraseological profile of texts, specialised corpora, and general reference corpora, and that concgram analysis impacts the learning and teaching of vocabulary in language syllabi (p. 250). One of the PLEC departmental sub-corpora, essays from students of the Department of Applied Biology & Chemical Technology (ABCT), was analysed for bi-grams, using Concgram 1.0 (Greaves, 2009). A total of 30,638 bi-grams with a frequency of two or more occurrences were found.

Table 5.1.2.1: Top 60 Bi-grams in ABCT essays

Rank	Bigram		Frequency	Rank	Bigram		Frequency
1	of	the	1642	31	for	the	314
2	the	to	1093	32	recycling	to	312
3	and	the	973	33	a	to	311
4	In	the	969	34	As	the	303
5	is	the	726	35	smokers	the	303
6	recycling	the	670	36	a	In	299
7	and	of	597	37	on	the	295
8	smoking	the	527	38	a	is	292
9	In	of	508	39	and	waste	289
10	is	of	490	40	Kong	the	276
11	the	waste	469	41	smoke	the	274
12	of	recycling	446	42	In	restaurants	271
13	of	to	432	43	recycling	waste	271
14	the	that	415	44	In	Kong	264
15	a	the	407	45	Hong	in	259
16	of	waste	397	46	Hong	the	258
17	a	of	383	47	not	the	258
18	In	to	382	48	and	is	251
19	and	to	381	49	It	is	249
20	In	smoking	375	50	restaurants	smoking	245
21	and	in	369	51	a	recycling	227
22	are	the	360	52	In	recycling	223
23	restaurants	the	357	53	the	will	223
24	It	the	354	54	of	smoking	222
25	In	is	351	55	government	the	221
26	be	the	346	56	banning	the	218
27	is	to	344	57	restaurant	the	217
28	can	the	343	58	and	recycling	215
29	Hong	Kong	333	59	of	that	214
30	is	recycling	326	60	smoking	to	214

It should be noted that the order of the words in the table above may not be the same as in the original text, because Concgram finds pairs regardless of order. Thus examples 45 and 46 include texts that are more likely to have *In Hong* and *The Hong* than *Hong in*. Example 44, *In Kong*, shows that *in* frequently occurs near *Kong*, regardless of intervening words like *In Hong Kong*. The advantage of this is that collocation is still recognised despite intervening words.

An investigation of n-grams in the PLEC corpus highlighted a special feature of the corpus. The essays prompt contained a number of facts for them to describe and analyse in the essay. Some students therefore have very similar sentences using these facts, resulting in long n-grams. For example, the sentence "Breathing secondhand smoke increase the risk of lung cancer and heart disease by about 25%" was a feature of 12 essays, ranging from students whose essays were graded as B+ to those which got D+. As this n-gram is 15 words long, and similar n-grams were found about another common topic, student debt, it is hard to isolate n-grams that might be a common feature of Hong Kong undergraduate students' writing in general when compared to BAWE students, due to this interference from the essay prompts. Removing these n-grams would be difficult, because students may only be using short extracts such as *the risk of* more frequently than in the n-grams. This was checked, and although it was found 109 out of 159 uses of *the risk of* referred to cancer caused by smoking, the n-gram was used in other essays; e.g. "and the risk of making the economic downturn more serious in Hong Kong".

The tendency of Hong Kong students to re-use language from essay prompts is commented on by Milton (2001), who states that 'L2 students… parrot the lexis and grammar of the examination prompts more readily than the L1 students' (p. xix) because of their restricted vocabulary, and gives statistical evidence that 'About 4% of all words in a 500,000-word corpus of Hong Kong students' examination scripts consist of eight adjectives repeated from the examination prompts' (p. 14), in contrast to 0.5% by UK students. He also found the long n-grams, including a 27-word string that was repeated 20 times in the corpus. The reason for this, he believes, is that the students 'do not have access to the lexical networks that make an NS's lexicon productive, such as hyponymy, synonymy, antonymy, etc.' (p. 14).

Bi-grams for BAWE-EON essays were also extracted, and show a different pattern, as can be seen in Table 5.1.2.2 below. The only word that seems to be academic vocabulary is *for example*, at position 30. In fact, it is not until position 127 that

Table 5.1.2.2: Bi-grams for BAWE essays 1-60

Rank	Frequency	Bigram	Rank	Frequency	Bigram
1	6786	of the	31	442	that it
2	3738	in the	32	432	in this
3	2495	to the	33	426	due to
4	1911	and the	34	414	there are
5	1771	it is	35	407	they are
6	1558	p fnote	36	399	the way
7	1541	that the	37	394	the same
8	1529	fnote fnote	38	386	the most
9	1403	to be	39	377	has been
10	1252	as a	40	374	the first
11	1153	on the	41	368	of this
12	1060	for the	42	367	use of
13	1054	of a	43	360	the world
14	1031	with the	44	360	was a
15	996	by the	45	357	would be
16	928	as the	46	355	of their
17	917	can be	47	339	was the
18	898	is a	48	334	does not
19	839	from the	49	332	able to
20	819	such as	50	331	between the
21	752	in a	51	326	could be
22	693	this is	52	324	in which
23	675	at the	53	323	one of
24	644	is the	54	322	fnote the
25	618	there is	55	320	rather than
26	588	is not	56	319	the new
27	585	it was	57	319	way of
28	518	to a	58	316	in order
29	493	have been	59	315	with a
30	445	for example	60	312	that they

the bi-grams become more academic, such as bi-grams for description, contrast, interpretation, prioritisation and argumentation.

Table 5.1.2.3: Bi-grams for BAWE essays on archaeology 100-160

Rank	Bigram	Frequency	Rank	Bigram	Frequency
100	206	on a	131	173	in fact
101	205	fnote ibid	132	171	by a
102	205	have a	133	171	New York
103	203	during the	134	171	such a
104	202	are not	135	169	based on
105	201	not be	136	169	he was
106	200	that a	137	169	the s
107	199	nature of	138	168	it can
108	199	not only	139	168	quote fnote
109	199	the poem	130	167	the people
110	195	this was	141	166	a more
111	194	and it	142	166	of its
112	193	however the	143	166	the social
113	192	in their	144	165	the play
114	191	about the	145	165	when the
115	188	as they	146	163	end of
116	187	suggests that	147	163	it would
117	185	which the	148	163	the only
118	184	the end	149	163	the state
119	182	over the	150	162	concept of
110	181	to do	151	162	from a
121	178	a new	152	162	the audience
122	178	had been	153	162	the main
123	178	that we	154	161	and so
124	177	is no	155	161	than the
125	176	all the	156	159	according to
126	176	if the	157	159	in terms
127	176	need to	158	159	the British
128	176	the war	159	158	and therefore
129	176	we are	160	158	pg fnote

Thus it appears that this finding has two conclusions, firstly that the PLEC essays tend to use subject-specific bi-grams more commonly, mainly because they are writing on the same topics and using the language from the essay prompts due to restricted vocabulary. Secondly the bi-gram distribution pattern accords with Cheng's (2012) observation that there is a continuum 'which one invariably finds in such frequency lists from the grammatically-rich co-occurrences at the top to the increasingly lexically rich as one moves down the list' (p. 97).

5.1.3 Disciplinary Variation

The finding that language is used differently in different disciplines in the BAWE corpus is detailed in Nesi and Gardner (2012, pp. 113-130). Regarding pronouns, for example, they found that the first-person pronouns *I* and *me* are used more frequently in Philosophy than in other disciplines, with 34.8 instances per 10,000 words, compared to Business at 6.8 and Engineering at 5.9, with a mean of 12.7 across all the essays in the corpus. There were more instances in humanities, which the authors attribute to value being placed on personal experience and the pronouns being used in framing moves such as *I have argued*, compared to less instances in science subjects, which they ascribe to disciplinary culture (pp. 115-7).

To compare this to PLEC, a business corpus (PLEC-BUSS) was made by combining the essays of students from the Accounting, Business, Management, and Logistics departments. An engineering corpus (PLEC-ENG) was compiled by combining the essays from students in the Building and Real Estate, Electrical Engineering, Electronic and Information Engineering, and Manufacturing departments. These were searched for *I* and *me*, using regular expressions to avoid matches with abbreviations such as *i.e.*, and the results are shown in the table below.

In Table 5.1.3.1 it can be seen that PLEC essays include more uses of first person pronouns than BAWE essays, regardless of discipline. This over-use may be because PLEC students, like ICLE students, were asked to write an argumentative essay in which personal opinions may be expected, while BAWE students wrote essays that Nesi and Gardner (2012, p. 98) categorise as exposition, discussion, challenge, factorial, consequential and commentary. First person singular pronouns do not appear in their examples for discussion, challenge, factorial, and commentary essays.

Table 5.1.3.1: Use of ‘I’ and ‘me’ in BAWE and PLEC Essays

Corpus of	Per 10,000 words			Frequency			Corpus
essays	I	Me	Total	I	Me	Total	size
All BAWE*			12.7			5092	4,010,103
BAWE-BUSS*			6.8			158	232,132
BAWE-ENG*			5.9			77	130,290
BAWE-EON	15.4	2.1	17.5	1010	139	1149	657,339
PLEC-EAP	34.0	2.5	36.6	2178	162	2340	640,013
PLEC-BUSS	26.8	2.7	29.6	322	33	355	120,032
PLEC-ENG	34.0	2.6	36.6	431	33	464	126,682

*BAWE figures from Nesi and Gardner (2012, pp. 113-4)

Nesi and Gardner also investigated keywords, both across all disciplines, and within disciplines. A keyword is a word that occurs ‘statistically more frequently in a small corpus than in a larger ‘reference’ corpus, relative to the total number of words in each corpus’ (Leedham, 2014, p. 42). Unfortunately, Nesi and Gardner give keywords only for disciplines that do not appear in the PLEC corpus, so no comparison can be made.

Nesi and Gardner (2012, p. 126) found the following keywords in the BAWE essays, including all years and all native languages. As can be seen in Table 5.1.3.2 below, all but two of the words, *kill* and *refuse*, are used more in BAWE-EON as compared to PLEC-EAP. This is probably a reflection of the fact that BAWE essays are on disciplinary topics. The disciplinary spread of the essays in BAWE is not even, as in the 1,238 essays 602 come from arts and humanities, 444 from social sciences, 127 from life sciences, and 65 from physical sciences, which could account for the high number of keywords that seem to related to history and social sciences, and low number of words related to life and physical sciences in the table below.

Table 5.1.3.2: Keywords in BAWE Essays

	Word	BAWE-EON		PLEC-EAP		Log-likelihood	Sig.	Log ratio	Use
		Freq	Per 100,000	Freq	Per 100,000				
Verbs	die	110	17.19	19	2.89	73.45	0.0000	2.57	Over
	fight	72	11.25	7	1.07	63.98	0.0000	3.40	Over
	fear	121	18.91	34	5.18	54.14	0.0000	1.87	Over
	assert	33	5.16	1	0.15	38.97	0.0000	5.08	Over
	born	69	10.78	24	3.65	23.93	0.0000	1.56	Over
	criticise	10	1.56	1	0.15	8.79	0.0039	3.36	Over
	deny	26	4.06	9	1.37	9.08	0.0043	1.57	Over
	kill	31	4.84	55	8.37	6.16	0.0064	-0.79	Under
	refuse	7	1.09	19	2.89	5.44	0.0137	-1.40	Under
	portray	32	5.00	0	0.00				Over
Adverbs	perhaps	403	62.97	35	5.33	373.10	0.000	3.56	Over
	entirely	100	15.62	1	0.15	131.46	0.000	6.68	Over
	merely	124	19.37	8	1.22	125.75	0.000	3.99	Over
	ever	120	18.75	12	1.83	105.48	0.000	3.36	Over
	ultimately	89	13.91	5	0.76	93.50	0.000	4.19	Over
	socially	59	9.22	1	0.15	74.57	0.000	5.92	Over
	seemingly	45	7.03	1	0.15	55.32	0.000	5.53	Over
	certainly	89	13.91	48	7.31	13.58	0.000	0.93	Over
	surely	44	6.87	25	3.81	5.82	0.016	0.85	Over
	essentially	60	9.37	0	0.00				Over
Adjectives	male	256	40.00	6	0.91	312.75	0.0000	5.45	Over
	ancient	175	27.34	2	0.30	228.12	0.0000	6.49	Over
	civil	151	23.59	1	0.15	202.71	0.0000	7.28	Over
	moral	205	32.03	34	5.18	140.41	0.0000	2.63	Over
	religious	128	20.00	16	2.44	102.18	0.0000	3.04	Over
	historical	89	13.91	9	1.37	77.88	0.0000	3.34	Over
	modern	384	60.00	215	32.73	52.96	0.0000	0.88	Over
	sexual	116	18.12	39	5.94	42.09	0.0000	1.61	Over
	liberal	131	20.47	50	7.61	39.76	0.0000	1.43	Over
	contemporary	91	14.22	0	0.00				Over
Nouns	war	621	97.03	6	0.91	818.01	0.0000	6.73	Over
	death	362	56.56	21	3.20	377.35	0.0000	4.15	Over
	century	400	62.50	35	5.33	369.36	0.0000	3.55	Over
	god	304	47.50	21	3.20	302.50	0.0000	3.89	Over
	truth	271	42.34	12	1.83	299.95	0.0000	4.54	Over
	character	221	34.53	3	0.46	284.56	0.0000	6.24	Over
	belief	159	24.84	5	0.76	186.74	0.0000	5.03	Over
	man	365	57.03	86	13.09	193.28	0.0000	2.12	Over
	religion	108	16.87	6	0.91	113.77	0.0000	4.21	Over
	society	830	129.68	465	70.78	114.26	0.0000	0.87	Over

There was also a large amount of variation between broad disciplines.

Table 5.1.3.3: Distribution of BAWE Keywords Across Broad Disciplines in BAWE-EON

	Word	Frequency per 100,000 words				Range	Range as % of Maximum	Variance	St. Dev.
		Arts and Humanities	Sciences						
			Social	Life	Physical				
Verbs	die	25	1	7	0	25	100	102	10
	fight	16	2	1	5	14	91	33	6
	fear	19	10	26	15	16	60	34	6
	assert	5	7	1	0	7	100	7	3
	born	15	3	4	3	12	82	25	5
	criticise	1	2	3	0	3	100	1	1
	deny	4	4	3	0	4	100	3	2
	kill	7	1	1	0	7	100	7	3
	refuse	0	2	3	3	3	92	1	1
	portray	6	2	3	3	4	68	3	2
Adverbs	perhaps	64	68	32	13	56	81	534	23
	entirely	19	8	6	10	13	69	25	5
	merely	17	24	18	5	19	79	47	7
	ever	16	22	10	28	18	64	44	7
	ultimately	15	11	9	3	13	84	22	5
	socially	7	15	7	0	15	100	28	5
	seemingly	6	11	1	8	10	87	12	3
	certainly	15	15	1	0	15	100	53	7
	surely	7	8	0	5	8	100	10	3
	essentially	8	14	4	10	9	68	12	3
Adjectives	male	30	79	12	3	76	97	869	29
	ancient	39	4	1	10	38	96	225	15
	civil	26	23	1	10	25	94	99	10
	moral	32	35	13	18	22	63	87	9
	religious	25	12	4	0	25	100	92	10
	historical	15	14	4	0	15	100	42	6
	modern	53	86	26	28	60	69	586	24
	sexual	19	19	10	0	19	100	61	8
	liberal	12	50	0	13	50	100	346	19
	contemporary	16	12	0	13	16	100	37	6
Nouns	war	133	37	6	5	128	96	2737	52
	death	80	6	19	10	74	93	897	30
	century	76	40	19	13	63	83	615	25
	god	68	6	10	3	66	96	730	27
	truth	60	9	1	5	59	98	579	24
	character	50	5	3	3	47	95	405	20
	belief	27	23	16	3	24	90	84	9
	man	73	29	15	10	63	86	619	25
	religion	16	25	3	0	25	100	101	10
	society	100	233	60	28	205	88	6078	78
	Mean:	31	24	9	7	34	89	407	14

Disciplinary variation in BAWE-EON can be seen in the Table 5.1.3.3 above, in which 35% of the words have zero occurrences in one of the broad disciplines, the largest range is 50 to zero for the word *liberal*, and the mean of the Range as a percentage of the maximum frequency of the words is 89%.

Nesi and Gardner (2012) conclude by stating that statistical and keyword analyses ‘provided clear evidence of the distinctive language of each genre family. They point to further disciplinary differences’ (p. 130). This is supported in the BAWE-EON corpus and by the over- and under-use in comparison with the PLEC corpus. Therefore there are implications for learning and teaching materials, to be discussed in the next chapter.

5.2 Organisation

There are many factors in the organisation of a text, such as cohesion, coherence, move structure, and whether the information is ordered chronologically or by some other principle. However, these are difficult to analyse with concordancing tools, so only one aspect of cohesion is addressed here: the use of connectors. This is an important area of concern, as students often have problems with them, partly due to problematic teaching materials, as described below.

5.2.1 Connectors

The use of connectors is, according to Conrad (2000, p. 550), register-specific. Due to the emphasis in academic writing on argumentation and logic, the use of linking adverbials such as *however* and *therefore* is more common than in such non-academic genres such as journalism.

Based on an analysis of the BAWE corpus, Leedham (2011) found that Chinese students make greater use of particular connectors. The connectors that Leedham identifies are: *however, therefore, besides, nowadays, in other words, meanwhile, and so on, what's more, on the other hand, nevertheless, last but not least, at that time, in the long run,* and *at the same time*. As can be seen in Table 5.2.1, in BAWE-EON 8,326 instances were found, or 1.2666%. In PLEC there were 3,843 instances, a percentage of 0.6005 and a significantly lower figure, thus Leedham's finding does not appear to apply to the PLEC student writers.

Breaking down these figures, Leedham (2011, p. 178) categorises *besides, what's more,* and *last but not least* as informal. *Besides* and *last but not least* are over-used in PLEC, but *what's more* appears in neither corpus. This overuse might be because students may have been taught word lists of connectors without differentiation according to formality, been taught inaccurate translations (Lee and Chen, 2009, p. 288), do not realise that one of the functions of *besides* is to indicate an afterthought, or are using it in a general, rather than academic, register.

Leedham categorises *however,* and *therefore* as 'negative' keywords, in that they are words which occur less often than would be expected by chance in comparison with a reference corpus (2011, p. 147). She gives the percentage from a reference corpus of 0.15% for *therefore* and 0.19% for *however*, (2011, p. 178). If BAWE-EON sub-corpus is regarded as a reference corpus, there are significantly less occurrences of both *therefore* and *however* in PLEC than BAWE-EON, thus confirming her expectation.

Table 5.2.1: Comparison of Leedham's Connectors in PLEC-EAP and BAWE-EON

Connector	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	Log ratio	Use in PLEC
	Freq.	Freq. / 100,000 words	Freq.	Freq. / 100,000 words				
however	1789	280	4745	722	1467.25	0.0000	-1.45	Under
therefore	917	143	2715	413	979.24	0.0000	-1.60	Under
nowadays	281	44	13	2	293.96	0.0000	4.40	Over
besides	239	37	50	8	129.40	0.0000	2.22	Over
nevertheless	62	10	191	29	72.46	0.0000	-1.66	Under
in other words	26	4	69	10	21.35	0.0000	-1.45	Under
on the other hand	310	48	243	37	6.45	0.0015	0.31	Over
at that time	17	3	41	6	10.89	0.0019	-1.31	Under
in the long run	18	3	33	5	4.89	0.0432	-0.91	Under
and so on	23	4	39	6	4.61	0.0525	-0.80	Under
at the same time	125	20	159	24	5.04	0.0726	-0.39	Under
meanwhile	27	4	28	4	0.05	0.9714	-0.09	Under
what's more	0	0	0	0	-			-
last but not least	9	1	0	0	-			-
Total	3843	600	8326	1266				

Regarding the over-use in PLEC-EAP of the lexical chunk *on the other hand*, Leedham (2014, p. 44) found that it was also over-used by Chinese students in the BAWE corpus compared to English students (her Eng123 corpus) in the same corpus. She suggests two possible reasons for this, firstly that 'For Chinese students, on the other hand may be frequently used as it is regarded as equivalent to a Mandarin expression literally meaning 'the other side of the problem' (一个问题的另一面 [yi ge wen ti de ling yi mian]) and is seen as having a more strongly contrastive meaning than the popular Eng123 connector *however*' (p. 45), and secondly that students select a longer lexical chunk in order to increase the word count of their writing.

That learners sometimes do not understand that *on the other hand* is a contrast connector is shown by the 5 instances in PLEC-EAP of the phrase *On the other hand, however, …*, which unnecessarily duplicates the contrast. An example is,

> "There are several advantages and disadvantages of constructing a second railway link to the mainland. It can solve the overcrowding problems at Lowu and create lots of job opportunities. *On the other hand, however*, it costs much money and cause environmental and social problems."

There is one example in BAWE-EON,

> "*On the other hand however* it is important to realize, that the Church itself would have been a relatively defenceless and extremely wealthy target."

Another possible cause for the misuse of *on the other hand* is errors in model answers. Leedham (2014, p. 84) gives a sample model from mainland China containing the following paragraph in a model answer on the reasons for cheating in examinations, "As students, we often take examinations at school, but sometimes we have too many examinations which are too difficult for us. *On the other hand,* some of us are lazy and don't work hard at their lessons." To me this is an error because both sentences are about negative causes, so there is no contrast between them. The second sentence should therefore have a connector of addition, not contrast. An example of this from PLEC-EAP is

> "Brown (1999) states another reason of students use credit cards is they can independent from their parents. They can have their own accounts and an independent status. Thus, most university students hold their own credit cards. (New paragraph) *On the other hand*, students use credit card can train their managing skill and responsibilities on their financial affairs."

The reasons of firstly independence from parents and secondly managing skill and responsibilities both seem to be positive aspects of students' use of credit cards.

Leedham (2014, p. 46) also points out that *therefore* and *however* are found more commonly at the start of sentences in her corpus of Chinese students' writing in BAWE, called Chi123, with 88% and 60% being at the start, whereas the figures are 65% and 31% in Eng123. Leedham gives a number of conjectures for this from the literature (pp. 46-7), and Paquot (2010) found that this practice was common across learner populations in ICLE, but in my opinion the sentence-initial positioning of these connectors may be because Chinese students have been taught to put the logical links at the start of sentences so they can be scanned easily by assessors, and so the logic is not lost by mixing it into the content. Therefore, this may be a case in which encouraging learners to imitate native-speaker norms may be counter-productive. Leedham highlights ‘the importance of avoiding over-generalization of findings’ (2011, p. 265). However, the encouragement of greater use of *therefore* and *however*, and the replacement of *besides* and *last but not least* with more formal versions such as *in addition,* and *finally* seem quite simple methods of increasing the logical organisation and formality of a piece of writing.

5.3 Conventions

This section covers referencing and citation, rhetorical questions, the first person plural, register, linguistic variation, stance and voice. These are all areas in which a writer can build a social relationship with the reader through text, for example by following academic norms regarding citation, expressing membership of an in-group of academics, displaying tentativity regarding findings in a way which follows group norms, and expressing their point of view appropriately and persuasively. The issue of inter-cultural differences also arises here, as it can cause issues with regard to the reader-writer relationship.

5.3.1 Referencing and Citation

One of the more important conventions in academic writing is citation. The corpora were searched for in-text citations in both bracketed number and author-date styles and in footnotes, and in PLEC 3,210 were found (0.4886%), and in the BAWE-EON essays 4,895 (0.7648%). Although more were found in the BAWE-EON essays, the difference is minor, as calculated by Ahmad’s (2005) weirdness formula, which divides the proportional use in the special corpus (in this case PLEC) by the proportional use in the general corpus (in this case the essays by English native speakers in BAWE-EON). The result of this calculation for citations and footnotes was 0.6388, which is less than the threshold for weirdness of 1.0. However, Hyland (2012) points out that the frequency and use of citations is discipline-dependent, so for mono-disciplinary classes, more specific analysis of citation practices in the field is advised.

5.3.2 Rhetorical Questions

The over-use of rhetorical questions is highlighted by Milton (2001), who comments that he found no rhetorical questions containing *you* in his UK student corpus. He explains that, in his opinion, such questions in an academic context signal doubt, but Hong Kong students use them to signal certainty. He concludes that such questions are inappropriate because 'The provocative stance that many Hong Kong students are encouraged to adopt is inappropriate in the context of open enquiry that is supposed to characterise academic, scientific, and most types of professional writing' (p. 13). In the PLEC-EAP corpus there are 13 questions that include the word *you*, but in BAWE-EON there are 22, although three of these seem to be from example survey questions. In CJA there were over 100, but most of these are from interviews or surveys being reported in the literature. Examples of these rhetorical questions are shown in the Figure 5.3.2, the CJA14 one being from an accounting review paper.

Figure 5.3.2: Rhetorical questions to 'you' in PLEC-EAP, BAWE-EON and CJA14

From PLEC-EAP:
Thus, what do you think about recycling now?
Do you like to eat in a restaurants with come smokers?
Do you agree that smoker will affect the other people?
Have you ever experience the following situation?
Do you agree students using or owning it?
Do you have credit card?
Would you like to use credit card?

From BAWE-EON:
Can you will this to become a universal law of nature?
How on earth are you going to set up something you don't know as the object of your search?
Have you not read of some such thing?
So what if I were to show you people wailing, screaming, pummelling the ground, covering themselves with earth?
For example how would you feel if I forgot your birthday?
Is God promising something on the condition that you behave in a certain way?

From CJA14:
But what if we told you only the amount each stock would report as its annual earnings per share in 12 months from now, and nothing else?

Reading the context around the BAWE-EON questions, it does seem that Milton is right and that the writers are expressing doubts, rather than to personalise issues as seems to be the case in many of the PLEC-EAP questions. However, when the LOCNESS corpus was searched, although no examples were found in the UK files, there were 26 in the USARG sub-corpus of American university students' essays, and these were a mixture of doubt questions such as "Scientists are inventing new methods to treat viral infections. Do you think they will be able to keep up with the rate of natural mutations of diseases?", which then goes on to talk about AIDs and the many other mutated diseases, implying that the answer to the question is at least doubtful and probably negative. Other questions suggest certainty, for example "How far would you go to be the best athlete in the world? Some people would go too far. Steroids and the use…"

There is also a degree of discipline-specificity, that in BAWE-D none of the questions come from life science, and over half are from arts and humanities; however, the situation was different in PLEC-D, where there were none from arts and humanities, which may be due to the low size that sub-corpus of 69,135 tokens taken from essays from only two departments, and half came from physical sciences, which has 269,973 tokens and nine departments.

Despite this, it seems that the question type that Milton did not find in his UK corpus can be found in BAWE-EON, CJA14 and LOCNESS. Suggestions for handling this issue are in the Discussion chapter below.

5.3.3 Register

Students' utilisation of the correct register in academic essays is an important component of their ability to write appropriately for members of the academic community. Gilquin and Paquot (2008) investigated this and found that students too often used language more suitable to spoken English in their essays. Their research was based on comparison of a number of corpora, the native speaker corpus being the academic and spoken components of the British National Corpus (BNC) as well as the LOCNESS corpus, and the learner corpus being 14 language-specific sub-corpora of ICLE2.0, including the Chinese one. However, Leedham (2014, p. 33) points out the differences between learners who learn primarily through spoken input such as European students, and those with more limited speaking practice such as Chinese students, and for the latter, this would possibly limit the amount of informal spoken language that they use in writing. This contradicts Milton's (2001) findings, as he identified characteristics of spoken English in the written Hong Kong interlanguage of school leavers, for example the over-use of personal pronouns, plural nouns and non-numeric quantifiers such as *some*, and the under-use of articles and prepositions.

To test the applicability of their findings to the comparison of PLEC and BAWE-EON in this study, the lexical items that Gilquin and Paquot cite as being unsuitable for written academic register, and the alternatives they suggest, were compared in four corpora: PLEC, BAWE-EON, three corpora of British English students' essays in LOCNESS (not the American ones) called LOCNESS-BRSUL123, and CJA14. The results are shown in the following tables.

Table 5.3.3.1 below contains lexical items that Gilquin and Paquot found to be under-used, and Table 5.3.3.2 below shows lexical items that they found to be over-used in learner writing due to the items being more suitable to spoken discourse. Items marked with an asterisk (*) are ones where the PLEC results differ from Gilquin and Paquot's. In the case of *it seems (that)*, PLEC students had probably been taught to use these expressions for hedging. Regarding *of course* and *And* (using 'And' at the start of a sentence), PLEC students had probably been informed that these were too informal for use in academic essays.

Table 5.3.3.1: Comparison of Lexical Items in Written and Spoken Registers in 4 Corpora

		PLEC-EAP		BAWE-EON		LOCNESS-BRSUL123		CJA14	
		Freq.	Freq./ 100,000	Freq.	Freq./ 100,000	Freq.	Freq./ 100,000	Freq.	Freq./ 100,000
Expected underuse in PLEC	it seems*	231	35.14	147	22.97	10	10.45	311	5.02
	likely	208	31.64	218	34.06	12	12.54	3192	51.48
	it seems that*	141	21.45	40	6.25	1	1.04	72	1.16
	perhaps	35	5.32	403	62.97	47	49.11	966	15.58
	apparently	10	1.52	25	3.91	4	4.18	206	3.32
	it is possible that	5	0.76	14	2.19	2	2.09	157	2.53
	surprisingly	5	0.76	11	1.72	2	2.09	161	2.60
	it is reasonable to*	4	0.61	3	0.47	0	0.00	36	0.58
	assumption	1	0.15	46	7.19	3	3.13	1060	17.10
	interestingly	1	0.15	16	2.50	1	1.04	259	4.18
	presumably	0	0.00	7	1.09	1	1.04	213	3.44
	it is worth noting that	0	0.00	5	0.78	0	0.00	35	0.56
	this article examines	0	0.00	0	0.00	0	0.00	13	0.21
	topics addressed will include	0	0.00	0	0.00	0	0.00	0	0.00
	Total	641	97.51	935	146.09	83	86.73	6681	107.76
	Mean	46	6.97	67	10.44	5.93	6.20	477	7.70

Table 5.3.3.2: Comparison of Lexical Items in Written and Spoken Registers in 4 Corpora

		PLEC-EAP		BAWE-EON		LOCNESS-BRSUL123		CJA14	
		Freq.	Freq./ 100,000	Freq.	Freq./ 100,000	Freq.	Freq./ 100,000	Freq.	Freq./ 100,000
Expected overuse in PLEC	because	1611	245.08	764	119.37	191	199.59	6216	100.26
	I think	287	43.66	53	8.28	9	9.40	146	2.35
	really	281	42.75	94	14.69	52	54.34	291	4.69
	besides	239	36.36	16	2.50	1	1.04	182	2.94
	thing	193	29.36	98	15.31	13	13.58	244	3.94
	like (for exemplification)	160	24.34	103	16.09	15	15.67	768	12.39
	maybe	68	10.34	41	6.41	4	4.18	75	1.21
	first of all	67	10.19	6	0.94	6	6.27	30	0.48
	certainly	48	7.30	89	13.91	12	12.54	0	0.00
	of course*	38	5.78	51	7.97	28	29.26	570	9.19
	Let us	32	4.87	19	2.97	3	3.13	274	4.42
	Let's	28	4.26	0	0.00	0	0.00	22	0.35
	absolutely	27	4.11	25	3.91	4	4.18	45	0.73
	So	23	3.50	18	2.81	2	2.09	71	1.15
	from my point of view	21	3.19	0	0.00	0	0.00	0	0.00
	definitely	18	2.74	8	1.25	1	1.04	52	0.84
	that is why	15	2.28	2	0.31	1	1.04	13	0.21
	I would like to talk about	12	1.83	0	0.00	0	0.00	0	0.00
	by the way	12	1.83	2	0.31	1	1.04	13	0.21
	thanks to	11	1.67	10	1.56	3	3.13	67	1.08
	And*	11	1.67	31	4.84	1	1.04	75	1.21
	I want to talk about	5	0.76	0	0.00	0	0.00	0	0.00
	this is why*	5	0.76	5	0.78	2	2.09	37	0.60
	look like*	5	0.76	5	0.78	0	0.00	55	0.89
	it seems to me	4	0.61	3	0.47	3	3.13	5	0.08
	I am going to talk about	2	0.30	0	0.00	0	0.00	0	0.00
	though.	2	0.30	3	0.47	0	0.00	9	0.15
	to my mind	0	0.00	2	0.31	0	0.00	3	0.05
	Total	3225	490.61	1448	226.25	352	367.84	9263	149.41
	Mean	115.18	17.52	51.71	8.08	12.57	13.14	330.82	5.34

Gilquin and Paquot suggest that use of spoken forms decreases with increases in proficiency (2008, p. 54), and that therefore there should be a cline in the statistics, with learner writing on one end, novice writing in the middle, and expert writing at the other end. For this these, the PLEC-EAP corpus forms the learner writing, BAWE-EON and LOCNESS-BRSUL123 form the novice writing, and CJA14 forms the expert writing. The authors give the examples of *maybe* and *I think*, showing how they are overused in learner writing, and the statistics in the table above confirm this, and also show that the same is true of *because, thing, like* (for exemplification)*, first of all, So* (at the start of a sentence)*, definitely, that is why,* and *by the way*. A cline with the opposite orientation, with the expert corpus showing high use and the learner corpus showing underuse, was seen with *it is possible that, surprisingly, assumption, interestingly,* and *presumably*. The exception was the lexical chunk *it seems*, which was overused by learners in comparison to the experts, as was *it seems that*, although the LOCNESS students used it more than the CJA14 writers. As mentioned above, the overuse of these two phrases could be because they had been taught as hedging devices. They also have the advantage of being grammatically easy-to-use, as *It seems* + adjective + *that* and *It seems that* can be added to the start of any sentence that needs to be hedged, and does not change part-of-speech in a way that other hedging devices do, so students do not have to make a choice; e.g. between *possible, possibly* and *possibility*.

The PLEC and BAWE results were tested for significance, and the results are shown in the Table 5.3.3.3 below. Overall, Gilquin and Paquot's findings seem to be mirrored in these statistics, with a few exceptions. Of these exceptions, there were only two instances of *to my mind* in BAWE-EON, which is very low and can probably be ignored. The adverb *certainly* is one that students have probably been taught to avoid in input on hedging. The use of *And* in the sentence initial position, although frowned upon as a connector, and misused all 11 times in PLEC, is used 75 times in CJA14, in expressions such as *And thus, And so,* and *And yet*, so student may have seen it used and not realised its distinct functions.

Table 5.3.3.3: Statistically significant register misuse in PLEC and BAWE-EON

	Word/Phrase	PLEC		BAWE		Log-likelihood	Sig.	Log ratio	Use
		Freq.	Freq. per 100,000	Freq.	Freq per 100,000				
Expected underuse in PLEC	perhaps	35	5.32	403	62.97	373.10	0.00	-3.56	Under
	it seems that*	141	21.45	40	6.25	57.06	0.00	1.78	Over
	assumption	1	0.15	46	7.19	56.69	0.00	-5.56	Under
	it seems*	231	35.14	147	22.97	16.65	0.00	0.61	Over
	interestingly	1	0.15	16	2.50	16.36	0.00	-4.04	Under
	apparently	10	1.52	25	3.91	7.05	0.01	-1.36	Under
	it is possible that	5	0.76	14	2.19	4.68	0.04	-1.52	Under
Expected overuse in PLEC	Let's	28	4.26	0	0.00	-	-	-	Over
	from my point of view	21	3.19	0	0.00	-	-	-	Over
	I would like to talk about	12	1.83	0	0.00	-	-	-	Over
	I want to talk about	5	0.76	0	0.00	-	-	-	Over
	I am going to talk about	2	0.30	0	0.00	-	-	-	Over
	to my mind*	0	0.00	2	0.31	-	-	-	Under
	because	1611	245.08	764	119.37	286.62	0.00	1.04	Over
	besides	239	36.36	16	2.50	228.02	0.00	3.86	Over
	I think	287	43.66	53	8.28	170.86	0.00	2.40	Over
	really	281	42.75	94	14.69	92.63	0.00	1.54	Over
	first of all	67	10.19	6	0.94	58.11	0.00	3.44	Over
	thing	193	29.36	98	15.31	29.10	0.00	0.94	Over
	like (For exemplification)	160	24.34	103	16.09	10.98	0.00	0.60	Over
	that is why	15	2.28	2	0.31	10.91	0.00	2.87	Over
	certainly*	48	7.30	89	13.91	13.58	0.00	-0.93	Under
	And*	11	1.67	31	4.84	10.46	0.00	-1.53	Under
	by the way	12	1.83	2	0.31	7.66	0.00	2.55	Over
	maybe	68	10.34	41	6.41	6.06	0.01	0.69	Over
	definitely	18	2.74	8	1.25	3.68	0.04	1.13	Over

Gilquin and Paquot give four possible reasons for this under- and overuse: influence of speech, L1 transfer, teaching-induced factors and developmental factors. Regarding influence of speech, they contrast English as a Second Language (ESL) learners whose language learning is primarily oral from authentic sources with English as a Foreign Language (EFL) learners who mainly learn from written sources, although with the help of a teacher who may be also a non-native speaker and have a limited knowledge of academic register. The ESL learners academic discourse is thus limited by their knowledge of English based on oral reception, whereas the EFL learners are hampered by the limited input they have access to (2008, p. 50). Although English is an official language in Hong Kong, most Hong Kong secondary school pupils are taught by non-native speakers, sometimes with the assistance of native speakers for speaking classes (Nunan, 2003, p. 599). In such situations students need specialist input at university by experienced academic writers.

The second possible reason that Gilquin and Paquot postulate for spoken forms in students' academic writing is L1 transfer, although the examples that they give are from French learners. An example from Cantonese is the expression *every coin has two sides*, meaning that there are advantages and disadvantages to every issue. This idiomatic expression and its variants are used 15 times in PLEC-EAP, as shown in the following concordance extract.

Figure 5.3.3: Concordance for 'coin' in PLEC-EAP

1 cut. Yet, the **other side of the coin** has not been examined thus the
2 really a good thing. Since **every coin has two sides**, cyber cafes also
3 to ask parents to help them. **A coin has two sides**. Credit cards can
4 game in those cyber cafes. **A coin is 2-sided**. There are positive
5 because of such ban. Meanwhile, **a coin have two sides**. The catering
6 would be highly unpopular. **A coin has two sides**. Developments have
7 a whole view of **both sides of a coin**, the advantages and disadvantage
8 There is the **other side of the coin**. Brown, A. N. (1999) states that
9 mother. **On the other side of the coin**, people think that illegal abort
10 young people to save their money. **Coin has two sides**, many young people
11 from Mainland China. **Just like a coin has two faces**. According to The
12 opportunities for local residents. **Coin has two sides**. There are two
13 everything **have two sides such as a coin**, using credit cards are no
14 to pay money immediately. **Every coin have two side**, although credit
15 **There are two side of the coin**, some people claim that users us

The third element in Gilquin and Paquot's explanation of the spoken-like nature of learner writing is teaching-induced factors. An example of this is teaching material that provides lists of lexical items that fulfil a function without differentiating between them in terms of the text types that they are appropriate for or whether their use is primarily spoken or written. Gilquin and Paquot (2008) describe this as the “pernicious influence of undifferentiated lists of connectors” (p. 53).

The final reason that they suggest is developmental factors, which are faced by both native and non-native language users as they learn a new genre. As novice academic writers, students are in the process of acquiring the rules and academic writing and lack knowledge of the range of alternative structures that they could utilise in their texts. Gilquin and Paquot analysed the LOCNESS corpus of British and American writers and found that they shared learners’ problems with register, including overuse of spoken expressions in writing. The table above shows a similar pattern, although in this thesis the LOCNESS corpus is limited to British students only, so that it parallels BAWE. Gilquin and Paquot concentrate on the expressions *maybe* and *I think*, and see a cline of use from the ICLE students to the BNC academic writers. This is mirrored in the first table in this section above, in which *maybe* has a frequency per 100,000 of 10 in PLEC-EAP, 6 in BAWE-EON, 4 in LOCNESS-BRSUL123 and 1 in CJA14. Similarly, for the phrase *I think*, the frequencies per 100,000 are 43 in PLEC-EAP, 8 for BAWE-EON, 9 for LOCNESS-BRSUL123 and 2 for CJA14.

Gilquin and Paquot conclude by pointing out that ‘it is often difficult to pinpoint which factor is responsible’ (p. 55). They go on to suggest further research using the BAWE corpus, after which academics ‘will then be ready for the final step, namely the preparation of appropriate remedial materials that will help learners overcome register-related problems’ (p. 55). The second table in this section above identifies which of the lexical items that they analyse are significantly different in frequency in PLEC-EAP and BAWE, and could be included in such materials.

5.3.4 Dimensions of Linguistic Variation in Register and Genre

As discussed above, register is an important factor in writing appropriately for a genre. A framework that explores register and genre was created by Biber (1988), based on a statistical analysis of linguistic features found in two corpora: the Lancaster-Oslo-Bergen (LOB) corpus and the London-Lund Corpus of Spoken English. He found that there were six dimensions, upon which PLEC and BAWE-EON are compared and contrasted below. The comparison was done using Nini's (2014) Multidimensional Analysis Tagger 1.3 (MAT) software, following the example in Crosthwaite (2016). This software uses the Stanford tagger to tag and then analyse 85 features of the text, ranging from information such as type-token ratio, to measures of the use of single lexical items, for example amplifiers such as *absolutely*, to clause-level constructs such as independent clause coordination. The software generates scores for Biber's dimensions from counts of these features, based on frequency per hundred tokens. The weightings for these are given in Nesi and Gardner (2012, pp. 269-271), who used Biber's dimensions to analyse the BAWE corpus. Some weightings are negative, leading to the negative scores below. The scores for all six dimensions are shown in the table below, and then are analysed individually in the subsequent paragraphs. The scores for each dimension are then used to state what type of text is being analysed, for example learned, scientific or general narrative exposition. Increasing or decreasing the frequency of various features will be recommended in order to align the text type with learned or scientific exposition, which is assumed to be more suitable for academic essays, rather than general narrative.

The BAWE corpus was analysed in terms of Biber's dimensions by Nesi and Gardner (2017) and Gardner, Nesi and Biber (2018), although they used a different set of dimensions, and compared a variety of genres rather than concentrating on essays. In general they found that ‘there is a distinctive register found in student writing that employs first person pronouns and ‘stance’ to- and that- clauses’ (2017). Regarding first-person pronouns, the MAT tagger scored the PLEC-EAP corpus with a mean score of 0.7, the same as CJA14. However the BAWE score was lower at 0.55. This reflects the finding in the section above on the first person plural, and may be because the PLEC students had been taught not to use these pronouns in their EAP subject. Stance is analysed in the next section of this chapter. Findings from Nesi and Gardner’s (2017) study are included in the analysis of separate dimensions below, where relevant.

Table 5.3.4.1: Multidimensional Analysis Tagger Scores for PLEC-EAP, BAWE-EON and CJA14

Dimension:	1	2	3	4	5	6	Register
	Involved vs. Informational production	Narrative vs. Non-Narrative Concerns	Explicit vs. Situation dependent reference	Overt Expression of Persuasion	Abstract vs. Non-Abstract Information	On-Line Informational Elaboration	
CJA14	-18.07	-3.21	6.79	-2.52	5.39	0.00	Learned exposition
BAWE-EON	-11.33	-1.67	6.74	-0.52	4.66	0.47	Scientific exposition
EAP A+	-1.99	-2.48	4.23	2.95	12.07	3.57	
EAP A	-8.62	-1.18	6.02	3.12	5.78	1.36	
EAP B+	-8	-1.76	6.7	2.93	5.55	1.05	
EAP B	-4.74	-1.43	6.9	2.86	5.13	1.07	General narrative exposition
EAP C+	-5.02	-1.61	6.85	2.71	4.61	0.86	
EAP C	-4	-1.61	6.06	2.11	4.33	0.38	
EAP D+	-4.96	-1.38	5.67	1.94	3.64	0.55	
EAP D	-5.45	-2.71	6.87	1.35	3.97	-0.2	
EAP F	-3.99	-3.01	4.91	-1.36	3.99	-0.38	
EAP Mean	-5.2	-1.91	6.02	2.07	5.45	0.92	
Spearman’s r	-0.536	-0.057	0.064	0.164	0.791	0.455	
p	0.089	0.868	0.853	0.631	0.004*	0.160	

* $p < 0.005$

Note: Spearman’s r and p are used to assess whether there is a cline in scores from the lowest PLEC EAP score of F to the highest CJA14 score, or vice versa.

The first dimension is involved vs. informational production, in which a high score indicates intimate interpersonal exchange of information, which is judged based on the numbers of verbs and pronouns, whereas a low score indicates scientific and learned exposition or narrative based on the numbers of nouns and adjectives. Biber (1988) states that in this dimension 'all academic sub-genres are characterised by the features of highly informational production' (p. 193). Gardner and Nesi (2017) found that essays have the lowest scores in this dimension, especially in arts and humanities, ranging from minus six to minus ten, which is similar to the PLEC-EAP and BAWE-EON means in the above table respectively.

In this dimension there seems to be a cline in the scores, although it is statistically not significant, from the PLEC-EAP F grade essays with the lowest scores, rising to the PLEC A grade essays. The PLEC A+ essay was a single essay, so the sample size is too small to draw conclusions from. The BAWE-EON essays and CJA14 articles had lower scores, reflecting their scientific and learned exposition. The scores regarding some of these components are given in the table below.

Examining the nominalizations, average word length, prepositional phrases, type-token ration and the attributive adjectives columns shows that the mean PLEC-EAP score was less than both the BAWE-EON score and the CJA14 score, and there was a cline in average word length and total prepositional phrases from the lowest PLEC-EAP grade score to the CJA14 score, suggesting that students might benefit from input in these areas.

Table 5.3.4.2: Dimension One Contributing Negative Factors

	Corpus	Nominalizations	Total other nouns	Average Word length	Total prepositional phrases	Type-token ratio	Attributive adjectives
CJA14		4.78	26.24	5.26	11.99	236	9.15
BAWE-EON		3.54	25.31	5.01	11.42	201	7.75
PLEC	EAP A+	2.94	26.68	5.17	10.5	201	6.51
	EAP A	3.26	25.51	4.98	11.1	200	6.11
	EAP B+	3.53	26.18	4.95	10.89	210	6.35
	EAP B	3.54	25.37	4.9	10.41	172	6.32
	EAP C+	3.34	25.47	4.84	9.87	201	6.39
	EAP C	2.88	26.33	4.78	9.68	185	6.1
	EAP D+	2.88	26.73	4.78	9.46	202	6.16
	EAP D	3.58	26.35	4.81	9.47	204	6.6
	EAP F	3.69	25.12	4.91	8.24	192	8.16
	EAP All	3.29	25.97	4.90	9.96	196	6.52
Spearman’s r		0.091	-0.045	0.825	0.964	0.282	0.200
p		0.790	0.894	0.002*	0.000*	0.401	0.555

* $p < 0.005$

The second dimension involved narrative vs. non-narrative concerns, as indicated by the number of past tenses and third person pronouns, which can be seen in the table below. Probably due to the fact that the corpora contained essays, not narratives, the scores for this dimension were low for PLEC-EAP and BAWE-EON, and even lower for CJA14.

Past tense use was lowest in PLEC-EAP, which may be because the essay topics involved current concerns. The Spearman’s rank correlation shows that there is a trend for greater use of the perfect aspect with PLEC-EAP grade. CJA14 and BAWE-EON have fewer present participial clauses than any grade in PLEC. These are clauses starting with an -ing form, followed by a preposition, determiner, pronoun or adverb (Nini, 2014, p. 22), for example “Being a student, …”. Overall, no strong recommendations for teaching can be made from this dimension, other than that already made for the perfect aspect in Section 4.3.2 above on Tenses.

Table 5.3.4.3: Dimension Two Positive Contributing Factors

Corpus		Past tense	Third person pronouns	Perfect aspect	Public verbs	Synthetic negation	Present participial clauses
CJA14		1.85	0.75	0.38	0.44	0.1	0.18
BAWE-EON		2.59	1.98	0.52	0.63	0.13	0.17
PLEC	EAP A+	1.26	1.05	0.63	0.84	0	0.21
	EAP A	0.92	1.98	0.51	0.77	0.06	0.38
	EAP B+	0.70	1.75	0.39	0.57	0.11	0.36
	EAP B	0.75	1.93	0.30	0.63	0.12	0.40
	EAP C+	0.76	2.18	0.27	0.56	0.12	0.38
	EAP C	0.85	2.38	0.25	0.58	0.12	0.36
	EAP D+	0.96	2.54	0.25	0.55	0.15	0.36
	EAP D	0.73	1.8	0.22	0.41	0.10	0.31
	EAP F	0.78	0.47	0.31	0.47	0.24	0.16
	EAP All	0.86	1.79	0.35	0.60	0.11	0.32
Spearman's r		0.573	-0.170	0.761	0.484	-0.484	-0.066
p		0.066	0.616	0.006*	0.131	0.131	0.847

* $p < 0.01$

The third dimension concerns explicit vs. situation-dependent reference. High scores indicate context-independence, such as in academic prose, with many nominalisations. Low scores indicate context-dependent language, such as sports broadcasts. Biber (1988, p. 193) states that all academic prose sub-genres have high scores on this dimension. In the dimension, PLEC-EAP averaged 3.02, BAWE-EON scored 6.74 and CJA scored 6.79. The positive factors in this dimension are the use of time, place and other adverbs, however, no significant patterns were observed in the scores for these. The negatively weighted factors are shown in the table below. Of these, comparing the scores for CJA14, BAWE-EON and PLEC-EAP All for nominalization indicate that it generally increases with ability, with PLEC-EAP averaging 3.29, BAWE-EON getting 3.54, and CJA getting 4.78. The Spearman's correlation for the different PLEC grades is low because there is no cline across PLEC-EAP scores. Therefore it seems that nominalisation should be encouraged in students' academic writing. In addition, pied-piping relative clauses, which contain a preposition followed by *who, whose* or *which,* such as *the way in which this happens,* seem to be much more common in CJA14 and BAWE-EON. A search of CJA14 for pied-piping relative clauses gave 8,142 matches in 714 out of the 760 texts, or an average of over 10 per text. The same search applied to BAWE-EON found 937 matches in 281 of the 330 texts, a mean of 2.8 per text. However, in PLEC-EAP, only 58 matches were found in 54 essays, with a maximum of one per essay. This mismatch in the frequency of use of this structure indicates that this type of relative clause may also be worthy of consideration for inclusion in teaching.

Table 5.3.4.4: Dimension Three Negative Contributing Factors

Corpus		WH relative clauses on object position	Pied-piping relative clauses #	WH relative clauses on subject position	Phrasal coordination	Nominalizations
CJA14		0.02	0.12	0.09	1.12	4.78
BAWE-EON		0.04	0.11	0.22	1.22	3.54
PLEC	EAP A+	0	0	0.42	0.21	2.94
	EAP A	0.01	0.01	0.28	0.9	3.26
	EAP B+	0.01	0.01	0.27	0.98	3.53
	EAP B	0.03	0.01	0.29	1.02	3.54
	EAP C+	0.03	0.01	0.31	1.03	3.34
	EAP C	0.03	0.01	0.25	1.01	2.88
	EAP D+	0.03	0	0.22	0.95	2.88
	EAP D	0.03	0.01	0.2	1.11	3.58
	EAP F	0.08	0	0.16	0.71	3.69
	EAP All	0.03	0.01	0.27	0.88	3.29
Spearman's r		-0.448	0.625	0.198	0.273	0.091
p		0.167	0.040*	0.560	0.417	0.790

* p < 0.05

# A pied-piping relative clause (e.g. *the manner in which he was told*) consists of any preposition followed by *who, whose* or *which*, and then a clause.

The fourth dimension is overt expression of persuasion, in which for high scoring language use the author's point of view is explicitly marked, for example containing hedges and modal verbs. Biber (1988, p. 194) explains that there is considerable variation among academic texts in this dimension. Surprisingly, PLEC-EAP scored more highly that the other two corpora on this. Analysing the contributing factors, the modal verb use for the possibility modals – *can, may, might,* and *could* was higher in PLEC-EAP, as were the predictive modals such as *would* and the necessity modals *ought* and *should*. This may be because the argumentative essays in PLEC are more likely to contain hedges containing *can, may, might*, and *could*, recommendations using *should,* and the effects of these recommendations using *would,* because the essay topics are on issues such as smoking and student finance. Examples from the corpus include 'It *may* expose youngsters with unwanted materials and affect their study', 'Therefore, students *should* have proper know-how and guidance so as to make them responsible and accountable for where the money is spent' and 'That *would* have a negative effect on their academic performance'.

Table 5.3.4.5: Dimension Four Contributing Factors

Corpus		Dimension 4: Overt Expression of Persuasion	Infinitives	Predictive modals - will, would, shall	Suasive verbs	Necessity modals - ought, should, must	Possibility modals - can, may, might, could
CJA14		-2.52	1.3	0.23	0.4	0.11	0.59
BAWE-EON		-0.52	1.68	0.43	0.41	0.18	0.68
PLEC	EAP A+	2.95	2.73	0.84	0.84	0	3.57
	EAP A	3.12	2.05	1.1	0.37	0.38	1.48
	EAP B+	2.93	1.97	1.01	0.34	0.37	1.64
	EAP B	2.86	2.12	1.03	0.28	0.36	1.68
	EAP C+	2.71	2.2	0.94	0.29	0.38	1.67
	EAP C	2.11	2.31	0.84	0.26	0.34	1.71
	EAP D+	1.94	2.19	0.96	0.22	0.34	1.64
	EAP D	1.35	2.37	0.9	0.22	0.24	1.47
	EAP F	-1.36	2.04	0.47	0.16	0.08	1.02
	EAP All	2.07	2.22	0.90	0.33	0.28	1.76
Spearman's r		0.164	-0.436	-0.193	0.952	-0.077	-0.143
p		0.631	0.180	0.569	0.000*	0.821	0.674

* $p < 0.001$

Table 5.5.4.5 above shows a cline in the use of 'suasive' verbs. These are verbs such as *agree, determine, insist, propose, recommend, require,* and *suggest* (Nini, 2014,

p. 29). However, the list also includes words that do not seem particularly academic, such as *beg, command, enjoin, pledge,* and *stipulate*, so attempting to raise the number of these in students work might be counter-productive.

A decrease in the scores related to student level for Dimension 4 in the BAWE corpus was also noticed by Nesi and Gardner (2017) in their analysis of the BAWE corpus using Biber's dimensions. Level one and two students' score was -1.4, rising to -1.5 in third year, and -2.0 for Masters' students. This is reflected in the BAWE-EON score of -0.52 and lower CJA14 score of -2.52.

Thus it seems that overt persuasion is not something that should be recommended for academic essay writing, and the PLEC students also hedged well, so no recommendations for additional teaching are made for this dimension.

The fifth dimension is abstract vs. non-abstract information. In this dimension high scoring texts contain technical, abstract or formal language containing many passive clauses and conjuncts such as *moreover*. Low scoring texts contain intimate interpersonal and informational interaction. Examining the overall PLEC-EAP score on this dimension, a cline can be seen with increasing scores from grades F to A+. However, the mean PLEC score is slightly more than both BAWE-EON and CJA14. Examining the contributing factors, it can be seen in the table below that PLEC-EAP students are using more conjuncts such as *moreover*, which accords with my personal experience and may be a result of teaching materials that contain undifferentiated lists of connectors. The statistics for agentless passive and by-passives show a clear cline from low use for the PLEC-EAP F grade essays to the A+ essay, and the BAWE-EON and CJA14 figures are above all the PLEC-EAP essays except the A+

one. There is also a cline in the use of past participial clauses e.g. "*Built in a single week*, the house would stand for fifty years". Therefore students can be advised to reduce the number of conjuncts that they use, and replace them with other cohesive devices such as thematic progression through theme and rheme (Dejica-Cartis and Cozma, 2013, p. 891).

Their awareness can also be raised regarding agentless passives, although care should be taken regarding genre, as Biber (1988, p. 194) found large differences in genre based on how technical they are. The use of past participial clauses is grammatically quite complicated in my opinion, and more common structures could communicate the same information, for example "Built in a single week, the house would stand for fifty years" could be replaced by "Although the house was built in a single week, it would stand for fifty years." Therefore academics would need to consider the teachability of this feature.

Table 5.3.4.6: Dimension Five Negative Contributing Factors

Corpus		Dimension5 Abstract vs. Non-Abstract Information	Conjuncts e.g. moreover	Agentless passives	By-passives	Past participial clauses	Other adverbial subordinators
CJA14		5.39	0.72	1.31	0.18	0.09	0.19
BAWE-EON		4.66	0.72	1.30	0.17	0.08	0.16
PLEC	EAP A+	12.07	1.89	2.31	0.21	0	0
	EAP A	5.78	0.87	1.22	0.1	0.06	0.18
	EAP B+	5.55	0.89	1.00	0.13	0.04	0.21
	EAP B	5.13	0.88	1.00	0.13	0.03	0.17
	EAP C+	4.61	0.87	0.87	0.12	0.02	0.15
	EAP C	4.33	0.84	0.76	0.08	0.03	0.15
	EAP D+	3.64	0.80	0.67	0.09	0.03	0.12
	EAP D	3.97	0.84	0.67	0.06	0	0.14
	EAP F	3.99	0.78	0.39	0.16	0	0.24
	EAP All	5.45	0.96	0.99	0.12	0.02	0.15
Spearman's r		0.791	0.043	0.968	0.652	0.700	0.107
p		0.004**	0.900	0.000***	0.030*	0.016*	0.755

Note on p-level: *** $p < 0.001$; **$p < 0.005$; * $p < 0.05$

The final dimension is on-line informational elaboration. High scores on this variable indicate that the text is informational in nature but produced under certain time constraints, as for example in speeches. Due to the lack of time limit for both BAWE-EON and CJA14 writing, the scores were very low. A high score on this dimension also means that the text presents many postmodifications of noun phrases. Regarding PLEC-EAP only, there was a distinct cline from F grade essays with the lowest scores, to the A+ grade essay with the highest, with a Spearman's rank correlation of 0.933 and a p of zero. Based on this, no recommendation can be made for teaching, only for further research using a corpus of non-timed student essays.

Table 5.3.4.7: Dimension Six Contributing Factors

Corpus		Dimension 6 On-Line Informational Elaboration	Dimension 6 (PLEC-EAP only)	*That* adjective complements	*That* verb complements	*That* relative clauses on subject position
CJA14		0.00		0.03	0.36	0.27
BAWE-EON		0.47		0.06	0.50	0.21
PLEC	EAP A+	3.57	3.57	0.21	1.47	0
	EAP A	1.36	1.36	0.03	0.76	0.05
	EAP B+	1.05	1.05	0.06	0.67	0.04
	EAP B	1.07	1.07	0.06	0.66	0.07
	EAP C+	0.86	0.86	0.05	0.65	0.08
	EAP C	0.38	0.38	0.04	0.6	0.07
	EAP D+	0.55	0.55	0.04	0.62	0.06
	EAP D	-0.2	-0.2	0.05	0.5	0.05
	EAP F	-0.38	-0.38	0.08	0.47	0.08
	EAP All	0.92		0.07	0.71	0.06
Spearman's r		0.455	0.933	-0.089	0.202	0.134
p		0.160	0	0.796	0.551	0.694

Regarding factors that influence this dimension, Biber (1988, p. 195) suggests that *that* complements to verbs and adjectives, as well as relative clauses, are used to mark information that cannot be carefully planned and integrated. For *that* verb complements and relative clauses the figures for BAWE-EON and CJA14 are lower than the PLEC-EAP mean, and for that adjective complements the same was true for CJA14, which may indicate more careful planning and integration was possible than for the PLEC students' timed essays.

Putting all the data about the six dimensions together, the scores are utilised by the MAT software to calculate the closest text type. For the PLEC-EAP essays of grade B and below, including the mean of all the PLEC-EAP essays, the text type identified was General Narrative Exposition, the calculation for which was based on the low scores for Dimension 1, Involved vs. Informational production, and high scores on Dimension 2, Narrative vs. Non-Narrative Concerns. PLEC-EAP B+ to A+ grade essays and the BAWE-EON corpus were categorised as Scientific Exposition, based on a low score for Dimension 1 and high scores on Dimensions 3 and 5: Explicit vs. Situation dependent reference and Abstract vs. Non-Abstract Information. The CJA14 corpus was categorised as Learned Exposition, based on the same dimensions as Scientific Exposition. Crosthwaite (2016) identified nominalisation as the deciding feature, stating that "the more nominalisation occurs, the closer a text type will match 'learned exposition' and not any other text type" (p. 8).

Overall, the comparison of the three corpora using the Multidimensional Analysis Tagger seems to confirm that the texts generally conform to Biber's dimensional framework, and suggest areas such as passive voice in which students could bring their writing more closely into accordance with academic prose norms and learned exposition.

5.3.5 Stance and Voice

An important feature of the social interaction between students and teachers that takes place via the medium of the essay is how learners present their stance and voice, as an essay is a persuasive communication act (Jiang, 2017, pp. 85-6). Although there is ambiguity in the definitions of stance and voice in the literature (Hyland and Guinda, 2012, p. 1), for the purposes of this research, definitions are taken from the range of definitions in Jiang, in which *stance* refers to "the ways writers express their personal views" (p. 86), while *voice* refers to the argumentative techniques that the writer uses in consideration of how their points might be received by the reader, including "taking into account their likely objections, background knowledge and rhetorical expectations" (p. 87).

As an example of how voice and stance are expressed, Jiang (2017) analysed the construction "noun + *that"*, for instance in phrases such as *the advantage that, in the mistaken belief that,* and *my suggestion that*. To find these phrases he searched the academic sub-section of the British National Corpus (BNC) and randomly selected a range of ten academic articles from each of a variety of disciplines. He categorised his results into entity, attribute, and relation nouns. The entity nouns had sub-categories of object, event, discourse and cognition, such as *report, fact, claim* and *view that*. Attribute nouns had sub-sets of quality, manner and status such as *danger, possibility* and *extent that*, while relation nouns covered differences as well as cause and effect, such as *grounds, result* and *reason that*. There were also nouns of suggestion, concealed stance, intertextual relations, and authoritarian nouns, for example, *fact, demonstration* and *insistence that*.

For this thesis, the nouns that Jiang identified were searched for in tagged versions of the corpora to ensure that the words which can be different parts of speech matched only with their noun forms. As can be seen in the tables below, major and significant differences were found between the writing in the PLEC-EAP and BAWE-EON corpora, with the latter containing about three times as many examples as the former, with the normalised frequency for PLEC-EAP being 49 per hundred thousand, and for BAWE-EON 143 per hundred thousand. This distribution was reflected in the suggestion, entity, attribute, and relation nouns, but there were no examples in PLEC-EAP for the authoritarian nouns. This may be, as Hyland (2008b) points out, a preference for impersonality by Hong Kong students found in other studies, which seems to result from both educational experiences and cultural preferences for a conciliatory, non-interventionist stance (p. 19).

In the paragraphs below there is an analysis of each of Jiang's categories, including quotations from the corpora in which the "noun + that" phrases are highlighted in italics. Phrases that did not occur in either corpus are omitted. There is also a comparison of disciplinary variation, as Jiang (2017) states that there is "considerable variation in the way that it (the 'noun + *that'* structure) is used to build knowledge across different disciplines" (p. 85).

Table 5.3.5.1 below shows the first category 'entity' and its first sub-category 'objects'. These objects refer to texts such as reports and studies, and the phrases demonstrate the stance and voice of the writer. As can be seen in the table, there were few uses of these phrases, which is similar to Jiang's finding of 0.04 uses per 100,000 words in his corpus, forming the lowest-used category.

The nouns were used in sentences such as “A *study that purports to* deal with social structure.... inevitably will reveal that the organisation or community is not all it claims to be”, taken from the BAWE-EON Social Science discipline. The writer’s stance can be seen in the verb *purports*, demonstrating doubt about the study. Instances in PLEC occurred across all disciplines, as did those in BAWE-EON, but references to studies were more common in arts and humanities in the latter. From the PLEC arts and humanities discipline an example is “Repace (2001, p. 24) *suggests in his study that* the prohibition of smoking in bars and taverns improves worker’s health.” Italics are inserted here for the reader’s easy identification of language usage examples in quotations from the corpora. None of the PLEC entity “noun + *that*” constructions contained an overt stance.

Table 5.3.5.1: Stance and Voice Noun + ‘that’ Constructions of Entity: Objects

Phrase	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	Log ratio	Use in PLEC
	PLEC Freq.	Freq. / 100,000	BAWE Freq.	Freq. / 100,000				
report that	2	0.30	0	0.00		-	-	Over
study that	3	0.46	3	0.47	0.00	0.97	-0.04	Over
studies that	0	0.00	4	0.62		-	-	Under
Total	5	0.76	7	1.09				

Note: the Fisher exact test result for ‘study that’ is p-value = 1

The next sub-category in Jiang’s analysis of ‘entity’ phrases is ‘events’. As can be seen in the Table 5.3.5.2 below, PLEC writers under-used these phrases in comparison to the BAWE-EON writers by a ratio of about 1:5, especially for the phrases “fact that” and “process that”.

The most frequently-used phrase is *fact(s) that*, which accords with Biber’s finding in Nesi and Gardner’s (2017) analysis of Biber’s dimensions in BAWE.

BAWE-EON writers expressed stance in sentences such as "Pearson *bemoans the fact that* most ancient funerary rites seem to be archaeologically invisible", "This *supports the evidence that* medieval peasants lived in agricultural settlements", and "However, *it is this very process that* can give rise to some of the most detrimental aspects of globalisation." PLEC students struggled with appropriate voice, for example in "It is *an incontrovertible fact that* smoking is harmful to one's health", which, while probably true, could be less strident. They also struggled with grammar, for example with the uncountability of *evidence* in "The study of Repace gave us *an evidence* that respiratory health can be improve if the environment is smoke-free" and "It is because from *those evidences that* mentioned before, abortion is against moral." Regarding disciplinary variation, the PLEC phrases ranged from 7 to 17 per hundred thousand words, with arts and humanities having least, but for BAWE-EON arts and humanities and social science had most, with over 50 per hundred thousand words, compared to about 23 per hundred thousand for life and physical sciences. This shows the discipline variability in both corpora.

Table 5.3.5.2: Stance and Voice Noun + 'that' Constructions of Entity: Events

Phrase	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p level	Fisher exact p-value	Log ratio	Use in PLEC
	PLEC Freq.	Freq. / 100,000	BAWE Freq.	Freq. / 100,000						
fact that	50	7.61	278	43.44	181	0.000	***	5.82×10^{-41}	-2.51	Under
facts that	0	0.00	4	0.62	-	-	-	-	-	Under
evidence that	20	3.04	29	4.53	2	0.167	-	0.195	-0.57	Under
change that	0	0.00	4	0.62	-	-	-	-	-	Under
changes that	0	0.00	9	1.41	-	-	-	-	-	Under
process that	1	0.15	11	1.72	10	0.002	**	0.003	-3.50	Under
processes that	0	0.00	4	0.62	-	-	-	-	-	Under
Total	71	10.80	339	52.97						

The next of Jiang's 'entity' sub-categories is 'discourse'. As can be seen in Table 5.5.5.3 below, the frequency per hundred thousand words is twice as much for BAWE-EON as for PLEC-EAP, and significantly more for *claims that* and *conclusions that*. The phrase *arguments that* is over-used in PLEC-EAP.

Regarding examples of use of these phrases, BAWE-EON students sometimes used them to express tentativity in stance when evaluating arguments, for example "If one accepts Jacques Derrida's *conclusion that* everything our minds have access is seen to be text, this theory has absolutely profound implications" and "Thus, what this discussion has entailed so far is that existence is a rudimentary predicate, but a predicate nonetheless. This is because it hasn't refuted *the claim that* existence possesses the essence of predication".

Table 5.3.5.3: Stance and Voice Noun + 'that' Constructions of Entity: Discourse

Phrase	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p lvl	Fisher exact p-value	Log ratio	Use
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
claim that	6	0.91	23	3.59	11.09	0.00	**	0.001	-1.98	Under
claims that	4	0.61	5	0.78	0.14	0.77		0.750	-0.36	Under
argument that	17	2.59	31	4.84	4.53	0.05		0.042	-0.91	Under
arguments that	9	1.37	3	0.47	2.98	0.07		0.146	1.55	Over
conclusion that	10	1.52	26	4.06	7.80	0.01	*	0.007	-1.42	Under
conclusions that	0	0.00	1	0.16	1.41	-		-	-	Under
suggestion that	2	0.30	6	0.94	2.20	0.16		0.174	-1.62	Under
suggestions that	0	0.00	4	0.62	5.65	-		-	-	Under
guarantee that	1	0.15	4	0.62	2.01	0.17		0.212	-2.04	Under
Total	49	7.45	103	16.09						

Note on p level: ** denotes p < .0001; * denotes p <= .01

However, PLEC students had some difficulty with expressing their stance, for example "It is *hardly to make a conclusion that* government should ban smoking on all restaurants or not" and "Evaluating the arguments for and against the importing professionals, *lend* me to draw the *conclusion that* Hong Kong should.” There is a need to teach students the collocates, and parts of speech of these collocates, of structures such as those marked in italics above.

There was little disciplinary variation in PLEC, with 5.8 to 9.6 instances per hundred thousand words, in contrast to BAWE, in which there were 25 per hundred thousand for social science, but none for physical science.

The next sub-category of ‘entity’ in Jiang’s categorisation is ‘cognition’. Again, the use of “noun + *that*” structures is about three times greater in BAWE-EON than PLEC-EAP.

BAWE-EON writers used significantly more of *view(s) that, idea(s) that* and *belief that*, while PLEC students used more *doubt that*. This was investigated further, and *no doubt that* occurred 80 times in PLEC-EAP, but only 12 times in BAWE-EON, giving a log-likelihood of 58 and p of 0. The structure *It is no doubt that* had 38 instances or 6 per hundred thousand words, for example “*It is no doubt that* the existing of cyber cafes has benefits and drawbacks to the society that I have discuss in my essay.” In BAWE-EON there were 10 instances of *there is no doubt that*, or about 1.5 per hundred thousand words, an example of which demonstrates a cautious voice in its concession to possible opposition from the reader, “Although some may feel that the genetic engineering carried out was unethical and possibly immoral, *there is no doubt that* it has advanced the learning and research on these issues”.

Regarding disciplinary variation of cognition constructions, in BAWE-EON, use in social science predominated, with 63 instances per hundred thousand words, and about 30-36 for the other disciplines. In PLEC social science was also top, with 19 per hundred thousand, compared to 8 to 15 for the others.

Table 5.3.5.4: Stance and Voice Noun + 'that' Constructions of Entity: Cognition

Phrase	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p level	Fisher exact p-value	Log ratio	Use in PLEC
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
view that	5	0.76	52	8.12	46.40	0.000	***	2.2 x 10^{-11}	-3.42	Under
views that	1	0.15	8	1.25	6.39	0.012	*	0.019	-3.04	Under
hypothesis that	0	0.00	7	1.09	-	-		-	-	Under
assumption that	0	0.00	29	4.53	-	-		-	-	Under
assumptions that	0	0.00	2	0.31	-	-		-	-	Under
idea that	8	1.22	97	15.16	91.39	0.000	***	4.2 x 10^{-21}	-3.64	Under
ideas that	2	0.30	13	2.03	9.31	0.002	**	0.003	-2.74	Under
belief that	1	0.15	52	8.12	64.92	0.000	***	3.2 x 10^{-15}	-5.74	Under
beliefs that	0	0.00	10	1.56	-	-		-	-	Under
doubt that	86	13.08	18	2.81	46.54	0.000	***	1.7 x 10^{-11}	2.22	Over
Total	103	15.67	288	45.00						

Note on p level: *** denotes p <.00001; ** denotes p < .0001; * denotes p <= .01

Jiang's next category is 'attribute', and the first sub-category is 'quality'. In contrast to the categories above, in this case the normalised frequency of the instances in the BAWE-EON corpus is slightly lower than the PLEC-EAP corpus, and three of the phrases are over-used in PLEC, as can be seen in the table below.

An example of the BAWE-EON social science students writing shows concern with research methods, "Either way, these points illustrate *the difficulties that* researchers face concerning access." In PLEC the phrase *advantage(s) that* was most popular, but some students had problems using it to summarise in their conclusion, for instance, "As *the advantage that* has point out in this essay, the students sometimes is using credit cards to buy something for educational or academic purpose".

Table 5.3.5.5: Stance and Voice Noun + 'that' Constructions of Attribute: Quality

Phrase	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p level	Fisher exact p-value	Log ratio	Use
	PLEC Freq.	Freq. / 100,000	BAWE Freq.	Freq. / 100,000						
danger that	0	0.00	2	0.31	-	-		-	-	Under
risk that	4	0.61	2	0.31	0.63	0.428		0.687	0.96	Over
advantage that	7	1.06	4	0.62	0.75	0.386		0.549	0.77	Over
advantages that	8	1.22	2	0.31	3.70	0.055		0.109	1.96	Over
difficulties that	0	0.00	2	0.00	-	-		-	-	Under
value that	0	0.00	2	0.31	-	-		-	-	Under
values that	0	0.00	4	0.62	-	-		-	-	Under
Total	19	2.89	18	2.50						

Regarding disciplinary variation, in BAWE-EON the figures stretched from 3.3 to 1.5 per hundred thousand for social and life sciences respectively, and for PLEC 3.3 to 2.2 for physical and life sciences in that order. So, although there are variations, the overall frequencies are so low that nothing much should be read into them.

Jiang's next sub-category in the category 'attribute' is 'status' nouns. Use in BAWE-EON is over 50% higher than in PLEC-EAP and *possibility that* is used more in BAWE-EON, but, as can be seen in the table below, *trend that* is significantly over-used and *choice that* is possibly over-used in PLEC-EAP.

Among the BAWE-EON students, one used it to criticise the arguments of an author, "He also *ignores the possibility that* rather than being a revolution in cinematic language, deep focus was more the result of excessive creative license on the part of people such as Greg Toland and Orson Welles." However, PLEC students had problems with hedging these phrases, for instance "It is an *inevitable trend that* recycling is replacing the role of landfilling and burning as a method of waste management", and even if they hedge, such as in "As it may decrease the choice that smokers can smoke" they are stating facts rather than expressing voice or stance.

They also have problems with grammatical agreement and missing determiners, such as “There *are possibility that* 21,500 job opportunities would be decreased after the total ban of smoking.”

There are striking differences in disciplinary variation in BAWE-EON, in which arts and humanities, and social sciences, have over 3 instances per hundred thousand words, but life and physical sciences have none. In PLEC, arts and humanities has 7, while the sciences have less than 2, instances per hundred thousand words.

Table 5.3.5.6: Stance and Voice Noun + ‘that’ Constructions of Attribute: Status

Phrase	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p level	Fisher exact p-value	Log ratio	Use in PLEC
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
possibility that	3	0.46	15	2.34	9.06	0.003	**	0.003	-2.36	Under
possibilities that	0	0.00	1	0.16	-	-		-	-	Under
probability that	0	0.00	3	0.47	-	-		-	-	Under
trend that	6	0.91	1	0.16	3.83	0.050		0.125	2.55	Over
choice that	4	0.61	1	0.16	1.85	0.174		0.375	1.96	Over
choices that	0	0.00	1	0.00	-	-		-	-	Under
abilities that	0	0.00	1	0.16	-	-		-	-	Under
Total	13	1.98	23	3.44						

Note on p level: ** denotes p < .01

The final sub-category of ‘attribute’ is ‘manner’. In this sub-category there is again three times greater use of the noun phrases in BAWE-EON than in PLEC-EAP, including significant under-use of *way that*, but significant over-use of *method that* in PLEC-EAP, as can be seen in Table 5.3.5.7 below.

BAWE-EON students sometimes use hedging to soften stance, for example in “Their suggestions are strongly related to the methods of textual analysis employed by Cultural Materialists…It is *arguably these methods that* can best help a reader make sense of The Waste Land.” They also sometimes use them to introduce specific examples of supporting evidence, for instance “There can also sometimes be *problems with scientific methods that* have to be overcome. There can be difficulties with radiocarbon dating, blood group analysis in lineage projects, and handling and labelling of mummies by later peoples."

PLEC students sometimes used these phrases inappropriately, for example in rhetorical questions, "Is it *time that* smoking affect your health actually?"

There was considerable disciplinary variation in BAWE-D, with a frequency of 20 per hundred thousand words for life sciences versus only 7 for physical sciences. In PLEC, 7 was the highest level, by social sciences, with arts and humanities lowest at 2.9 per hundred thousand words.

Students could be encouraged to use *extent that*, as it was not used in PLEC at all, and also to use *way that* more, for example in analogies, such as "However, if instead of resemblance it is interpreted as indicative representation, *in the same way that* smoke represents fire, not in the sense of resembling it, but in indicating it." The phrase *in the same way* is used 12 times, or 1.8 per hundred thousand words, in BAWE-EON.

Table 5.3.5.7: Stance and Voice Noun + 'that' Constructions of Attribute: Manner

Phrase	PLEC-EAP		BAWE-EON		Log-likelihood	Sig.	p level	Fisher exact p-value	Log ratio	Use
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
extent that	0	0.00	21	3.28	-	-		-	-	Under
way that	16	2.43	64	10.00	32.14	0.000	***	2.8×10^{-8}	-2.04	Under
ways that	3	0.46	5	0.78	0.56	0.454		0.502	-0.78	Under
time that	4	0.61	11	1.72	3.59	0.058		0.073	-1.50	Under
times that	1	0.15	0	0.00	-	-		-	-	Over
method that	7	1.06	1	0.16	4.90	0.027	*	0.070	2.77	Over
methods that	4	0.61	3	0.47	0.12	0.731		1	0.38	Over
Total	35	5.32	105	16.41						

Note on p level: *** denotes p <.00001; * denotes p < .05

The remaining categories in Jiang's framework do not have sub-categories. The first of these categories is 'relation'. As can be seen in Table 5.3.5.8 below, the frequency of phrases in this category between PLEC-EAP and BAWE-EON is much more similar than in many of the other categories, and the phrase *result(s) that* is used more in PLEC-EAP, whereas phrases that examine evidence, *findings, grounds, and reason(s)* are more common in BAWE-EON, although not at significant levels.

Table 5.3.5.8: Stance and Voice Noun + 'that' Constructions of Relation

Phrase	PLEC-EAP		BAWE-EON		Log-likeli-hood	Sig.	p level	Fisher exact p-value	Log ratio	Use
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
findings that	0	0.00	2	0.31	-	-		-	-	Under
grounds that	3	0.46	3	0.47	0.00	0.974		1	-0.04	Over
result that	6	0.91	1	0.16	3.83	0.050		0.125	2.55	Over
results that	1	0.15	1	0.16	0.00	0.985		1	-0.04	Over
reason that	10	1.52	14	2.19	0.78	0.377		0.419	-0.52	Under
reasons that	6	0.91	7	1.09	0.11	0.745		0.787	-0.26	Under
Total	26	3.96	28	4.37						

Examples of sentences from BAWE-EON that relate to evidence include "Their grasping hands and feet are also well suited to this purpose and it is *for this reason that* most anthropologists agree that the domination of an arboreal niche led to the development of many primate traits", and "The case of Boart Longyear *gives evidence to the findings that* employees tend to work harder in pursuing goals that they have helped set than those that have been assigned to them." PLEC-D sentences regarding results include "According to Sinclair (2000, p. 17), the Boston University School of Public Health have done a study, *it got a result that* there may be some new visitors to visit restaurants if the policy proposed and the non-smokers may go to restaurant for more time" and "And this situation may *lead to a result that* the unemployment rate of local people in Hong Kong would remain high and the situation of economic recession could not be improved."

Regarding disciplinary variation in the category of relation, in BAWE-D the highest frequency per hundred thousand was social science, at 9, with the others being less than 2. This could be because of a tendency in social sciences to cite multiple factors to support an argument due to the complexities of many social phenomena; e.g. "Thirdly, *there are many reasons that* lead to non-voting on the individual level." In PLEC-D, normalised frequencies ranged from 4.8 for physical sciences to 2.9 for arts and humanities. An example from physical sciences that isolates a single factor is "For *the reason that* they are not capable to repay their debts, as a result their parents may have to share their burden of debt."

Jiang next discusses a 'concealed' stance, which is one in which the writer is hidden, perhaps to show objectivity. Following the patterns in Jiang (2017, p. 100), these were searched for by the addition of *the* in front of the noun, for all of the nouns taken from Jiang's paper, using the part-of-speech tagged versions of the BAWE-D and PLEC-D tagged discipline corpora. This is not a perfect method, since the writer can be identified in other parts of the sentence, or surrounding sentences, so the concordance lines were scanned manually, and those that seemed to have an unconcealed source were discounted.

From Table 5.3.5.9 below it can be seen that for all disciplines except Physical Sciences the use in PLEC-D is significantly under that of BAWE-D. This indicates that BAWE-D writers are using many more concealed stance structures.

Table 5.3.5.9: Stance and Voice Noun + 'that' Constructions of Concealed Stance

	PLEC-D		BAWE-D		Log-likeli-hood	Sig.	p level	Fisher exact p-value	Log ratio	Use in PLEC
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
Arts and humanities	7	1.06	309	48.28	379.01	0.000	***	8.1 x 10^{-84}	-5.50	Under
Life sciences	22	3.35	36	5.62	3.80	0.051		0.065	-0.75	Under
Physical sciences	27	4.11	12	1.87	5.53	0.019	*	0.024	1.13	Over
Social sciences	22	3.35	161	25.16	122.98	0.000	***	3.5 x 10^{-28}	-2.91	Under

Note on p level: *** denotes p <.00001; * denotes p <.05

Examples of this include "Due to *the fact that* such a high status person was buried in such a manner without being a warrior suggests that there was more to 'Celtic' culture than the filthy 'barbarians' described by the invading Romans" and "Again this reflects *the idea that* they are going nowhere but also predicts that it is Billy who will fall behind first, as indeed he does." PLEC-D students gave examples such as "Added to this is *the fact that* children may absorb too many wrong ideas when they surfing some adult site", and "Of the various ways of protecting our environment, recycling is one of the methods that is commonly used nowadays." The BAWE-EON quotes seem to be interpretations of evidence, but the PLEC-D quotes seem to be giving evidence, and therefore seem more in need of citations.

The penultimate category examined by Jiang was intertextual relations, which is when a text refers to another text. A search was conducted for the nouns *demonstration, conviction, statement, claim, view, idea,* and *belief* + *that*, and the results were filtered manually for those which referred to the writer's own thoughts, for example because they contained a personal pronoun. There was significant under-use in PLEC-D in all disciplines, and these constructions were most common in social sciences in both corpora.

As can be seen in Table 5.3.5.10, although it might seem surprising that the number of intertextual relations phrases is quite low, this is probably because most intertextual relations is done with reporting verbs, rather than nouns. Examples of intertextual relations from BAWE-D include “On the other hand, the classical scholar A. B. Cook *expresses the view that* Minos the king and the character of the Minotaur are really the same individual in different outward appearances” and “Burket *has put forth the idea that* later ritual sacrifice was rooted in this early 'condition of man the hunter' when hunting and killing an animal was a spiritual, primal experience.” From PLEC-D examples include “In the same article, it also *implies the idea that* continue pregnancy is 10 times more risky than having an abortion” and “Some environmentalists *hold the view that* recycling is indispensable in Hong Kong.” In general the PLEC-D sentences contain frequently-used verbs such as *hold, get,* and *support*, whereas BAWE-D writers sometimes use less common verbs such as *expresses, emphasises, stems from, concur, undermine, subscribe to* and *put forth*. If learners are to implement the use of this type of “noun + *that*” intertextual relations construction, they need to use them more, and collocate them with a wider variety of verbs.

Table 5.3.5.10: Stance and Voice Noun + ‘that’ Constructions of Intertextual Relations

	PLEC-D		BAWE-D		Log-likeli-hood	Sig.	p level	Fisher exact p-value	Log ratio	Use in PLEC
	Freq.	Freq. / 100,000	Freq.	Freq. / 100,000						
Arts and humanities	2	2.9	163	38.8	211.44	0.00	***	6.3 x 10^{-47}	-6.39	Under
Life sciences	4	3.0	25	36.5	17.50	0.00	***	0.000	-2.68	Under
Physical sciences	13	4.8	8	20.4	1.07	0.30		0.384	0.66	Over
Social sciences	8	5.2	79	51.5	69.10	0.00	***	2.6 x 10^{-16}	-3.34	Under

Note on p level: *** denotes p <.00001

Jiang's final category is 'authoritarian', in which "writers give the floor to the voice of abstract entity such as institutions and authorities" (2017, p. 101). The corpora were searched for the noun phrases *acceptance, acknowledgment, acquiescence, admission, assent, assertion, authorization, concession, concurrence, condonance, confirmation, conjecture, compliance, corroboration, declaration, exhortation, hypothesis, injunction, insistence, instruction, postulate, prescription, presupposition, pronouncement, recognition, statement, supposition, theorem, theory*, and *verification that*. There were no matches in the PLEC-D corpus, but 71 in BAWE-D, comprising 47 for arts and humanities, 16 for social sciences, 4 for life sciences, and 3 for physical sciences, thus showing considerable disciplinary variation.

Examples from each discipline include "Locke's principle of individuation is the *theory that* ideas become general by separating from them the circumstances of time and place", "His technique of feeling the shape of the skull relied on the *theory that* the shape of the skull directly linked to the shape of the brain underneath it", "String Theory is a *theory that* does include gravity, and like the Standard Model, it is based around a model of elementary particles", and "The result of this may be a reduction in nationalism and its malign effects, a *theory that* Cobden raised during the Crimean War." From these examples it can be seen that *theory that* is the most common phrase, with 31 instances, or 4.6 per hundred thousand words. Other nouns and their occurrences were *acceptance* (1), *admission* (1), *assertion*(s) (10), *concession* (1), *hypothesis* (7), *insistence* (2), *recognition* (4), *statement*(s) (11), and *supposition* (2). Therefore teachers could consider raising students' awareness of how to cite authorities using (author's name)*'s theory/statement/hypothesis that*.

In conclusion, regarding stance and voice, from the examples above it can be seen that the BAWE-EON students are generally expressing stance and voice in a different manner, for example commenting on claims by authors and research methods, introducing examples, conceding to possible alternative points of view, and citing authorities, which contrast to the PLEC students' quite objective opinions which show little stance or voice. A possible reason for this is given by Hyland (2012), who states that 'rhetorical choices are not only influenced by the discipline to which the authors belong, but also by the specific cultural context in which texts are used' (p. 34). In one of Leedham's (2014) interviews of a Chinese student studying in the UK, the student said that "Chinese students are 'taught to be modest' and find it difficult to give their opinion" (p. 125). The attitude to critical thinking of Chinese learners studying in the UK was investigated by Durkin (2011), who found that in western cultures:

> emphasis in academic writing is laid on explicitness, where everything is stated very clearly and in a logical sequence. The Chinese students contrasted this with the indirect, inferential speech of Chinese cultures which is seen as more sensitive, and representing 'a higher level of communication' where 'everything is implicit', the hidden message is 'behind the language', and where the responsibility lies with the reader or listener to accurately interpret any ambiguities. (p. 284)

A second factor identified by Durkin (2011, p. 285) is that Chinese students in the UK see no long-term benefit from adopting these foreign models of critical thinking, and expect to return to a culture where critical skills may not be so acceptable. Students taking EAP subjects may also think this, especially if they know that, as is often the case in Hong Kong in their other subjects, the lecturer is Chinese.

This reluctance leading to a lack of stance and voice could be problematic, as Hyland (2008b) explains, because "the relative absence of their use in the student corpus suggests that these writers may be uncomfortable in explicitly aligning themselves with a particular evaluation or personally attesting to the weight they want to attribute to their claims" (p. 19) in a way that is expected in their academic writing.

Given that one of the current aims of the PLEC students' university is to promote critical thinking, stance and voice could be used by learners to draw attention to their abilities in this area, for example by referring to evidence and their appraisal of it. From the tables above it can be seen that learners could be advised to express their stance and voice more by the use of "noun + *that"* constructions, especially those with *fact, process, claim, conclusion, view, idea(s), belief, possibility, trend, way*, and *method that*. They should use *no doubt that* less. They can also use more concealed stance phrases when interpreting evidence. Students should take care to proof-read for agreement when using "noun + *that*" structures, for example with uncountables such as *evidence.* Teachers can also consider instruction on intertextual relations, and citing pronouncements by authorities.

5.4 Summary

This chapter examined content, organisation and conventions. The section on content showed that the PLEC essays tend to use subject-specific bi-grams more commonly, and that the bi-gram distribution pattern is normal. Disciplinary variation was evident in the over- and under-use in the comparison of the PLEC and BAWE corpora.

Regarding organisation, the PLEC students seem to over-use a number of informal connecting expressions, and their vocabulary training could include not only the meanings of expressions, but also information on their frequency and appropriacy.

Concerning genre conventions, based on research by Gilquin and Paquot (2008), it was found that students have a tendency to use spoken English lexical items in their academic writing, and items that significantly differ between PLEC-EAP and BAWE-EON were analysed, and can be used in teaching material.

In addition, the use of the MAT software (Nini, 2014) to analyse students' texts according to Biber's (1989) dimensions of register and genre revealed that lower level PLEC essays were categorised as General Narrative Exposition, while upper level PLEC essays and BAWE essays were labelled as Scientific Exposition. Students could be encouraged to include more Dimension 3 features of context-independent academic prose such as nominalizations.

The findings on stance and voice demonstrate that BAWE-D students express stance and voice in a more sophisticated manner, use language to build relationships with the reader more, and use a wider variety of "noun + *that*" phrases, with a wider range of collocations and more accurate grammar, compared to the PLEC-D students.

Relating these findings to the research questions of this thesis, the extent of the differences between the writing in BAWE-EON and PLEC is considerable and significant, and reveals some important aspects of student writing that can be incorporated into teaching materials. The next chapter discusses in more detail how this could be done using a research-informed approach.

Chapter Six: Discussion

This chapter examines how the findings from the previous chapter could be implemented in course design for students, by suggesting indirect and direct applications of corpus linguistics. Szudarski (2018) explains indirect and direct applications: in the former ‘corpora are used to inform the design and development of syllabuses, tests and teaching materials, while in the latter corpus data are used for data-driven learning (DDL); that is, hands-on activities in which learners themselves engage in corpus analysis’ (p. 96). Regarding materials produced for students, he emphasises that ‘it is of utmost importance that the language found in such materials reflects real-life communication that takes place in natural settings rather than be contrived solely for the purpose of covering a rigid teaching syllabus’ (p. 98), and thus material design should be informed by corpus-based research. This approach teaches students how and where to put words into sentences (Yoon and Hirvela, 2004, p. 260).

Not all corpus-based findings should be included in teaching materials, as the teaching context, such as the course aims and students’ abilities, are of primary importance. As Cook (1998) points out ‘computer corpora—while impressive and interesting records of certain aspects of language use—can never be more than a contribution to our understanding of effective language teaching’ (p. 58). Szudarski (2018) summarises teachers’ and students’ opinions of the usefulness of phrases from various corpus-based word lists, and concludes that ‘that the usefulness of words and phrases is a relative notion, with the ratings of what is pedagogically relevant likely to vary depending on specific teaching contexts’ (p. 102). These specific teaching contexts could be as specific as a single class or a single student learning autonomously. This is one of the reasons why this research does not involve the production and piloting of specific materials, instead being limited to detailing a range of possibilities of what could be included if found suitable to the context of the various EAP subjects in a variety of tertiary institutions in Hong Kong.

A second reason why this research does not do more than suggest possibilities is the 'lack of awareness of corpora, and, in some cases, resistance toward corpora from students, teachers and materials writers' observed by Römer (2011, p. 206).

To ameliorate this resistance, Römer suggests focussing attention on language teachers and their needs, which her survey revealed were better teaching materials and support for teachers when they are creating materials.[9]

In cases where the above corpora do not generate enough results, I have used Webcorp (Renouf, Kehoe & Banerjee, 2007) at http://www.webcorp.org.uk/live/ which uses the internet as its source of texts, and searches it in real time. An example of this is the word 'stuffs', which does not appear as a noun in written BAWE, the Brown corpus, or the BNC written corpus on Cobb's site, but Webcorp gives examples of it as a type of Indian textile, and therefore of possible relevance to the Fashion students at my university. However, a disadvantage of Webcorp is that the search results include texts by non-native speakers that often contain errors, although these can be filtered by looking at the URLs of the sources, and eliminating those that seem to be from non-native speakers. The root cause of this problem is that the

[9] Free online concordancers and corpora are suggested, examples of which are Cobb's Compleat Lexical Tutor at https://www.lextutor.ca , which includes not only a concordancer with a large number of corpora such as the BAWE written corpus, but also corpus-informed activities such as vocabulary profiles and a corpus builder that will combine text, HTML and Microsoft Word files into a single file, thus helping teachers build a corpus.

internet is not a corpus, because it does not follow Sinclair’s (2004a) definition regarding text being selected according to criteria and representing a language, as described in Chapter Three.

A third piece of online software is the Sketch Engine[10]. It has many features, such as a concordancer, part-of-speech tagger and word list generator (Kilgarriff et al., 2014).

If a language teacher has constructed their own corpus, as suggested by Aston (2000, p. 9) free software to analyse it such as a concordancer, word profiler and part-of-speech tagger are available from Laurence Anthony's Website http://www.laurenceanthony.net/software.html . However, the software license is free for personal use only.

Using the above software tools, language teachers can analyse academic writing, consider how the results of this analysis could be included in the curriculum, and construct teaching materials based on this analysis. The following section looks in more detail at how this can be done.

[10] Sketch Engine (https://www.sketchengine.eu) is commercial, although it has a 30-day free trial period.

6.1 Indirect Applications

The findings of this research described in the previous chapter can lead to indirect applications if they are used in materials design. The use of PLEC as a learner corpus, or teachers' use of corpora of their own students' work, is a valuable basis for this. Paquot (2010) states that 'By showing, in context, the types of infelicities EFL learners produce and the types of errors they make, as well as the items they tend to under- or overuse, learner corpora are the most valuable resources for designing EAP materials which address the specific problems that EFL learners encounter' (p. 206).

However, there is a need for caution in that, as Leedham (2014) warns, 'Since no corpus can be fully representative, I argue that corpus findings should be used cautiously as indicators of *tendencies* rather than definitive statements of language use' (p. 140).

A framework for creating corpus-designed activities has been proposed by Bennet (2010, pp. 18-20), and contains the following steps: ask a research question, determine the register on which your students are focused, select a corpus appropriate for the register (or compile authentic texts from that register), utilize a concordancing program for quantitative analysis, engage in qualitative analysis, create exercises for students, and engage students in a whole-language activity.

Indirect applications of the results of corpus-based investigations are discussed below, in the same order of analysis as Chapters Four and Five above.

6.1.1 Grammatical Features

An example of an indirect use of a concordance program for quantitative analysis, leading to a grammar exercise on the use of articles, is included in Johns (1994), who gives an example worksheet on the use of definite or zero articles in front of the word *industry*. It consists of a grammatical explanation based on qualitative analysis, and an exercise with gapped concordance lines for learners to complete. Johns comments that the lines should be carefully graded to allow the students to answer not only clear-cut instances, but also "fuzzier examples typical of authentic text" (p. 309). Johns also emphasises the importance of whole-language activity by suggesting a free sentence completion activity for students to improvise the second half of given sentences using the grammatical pattern being taught (p. 302).

A warning about the construction of learning materials based on concordance lines is given by Johns (1994, pp. 298-299). The danger is that the selection process necessary to choose lines for inclusion in activities might distort the evidence, whether because the teacher is selecting based on what they think ought to be in the data rather than analysing the data for patterns, or because they are selecting lines on good pedagogical criteria but they are biasing the sample in terms of frequency of the occurrence of certain features.

Johns gives three possible methods with which to address this problem. The first is the selection of a random sample of concordance lines, and while this might have the benefit of reflecting frequency, a large number of lines would be needed and the language might not be self-contained and not self-explanatory in terms of meaning, and thus be unsuited for teaching purposes. Johns does not use this method in his example materials. The second solution proposed is to analyse the lines for syntactic and communicative features and select lines suitable for teaching, although these might then display the frequency biases. The final method is to select data based on certain criteria, such as one example of every collocate of a lexical item, and to make that selection criteria known to students so that they do not assume that the frequencies are all equal. A possible enhancement to this, although not mentioned in Johns, would be to add frequency information to the samples, such as in what percentage of corpus texts the feature was found, or a frequency comparison between a pedagogical corpus and an expert corpus.

Research into the automatic generation of language learning exercises based on language use in corpora was carried out by Wilson (1997). However, she found a number of problems, including the need for corpora tagged for part of speech, lemma and sentence structure, the need for quality control of the texts to remove stylistically misused items, the need for texts graded to students' ability level, and the need for a large enough corpus that it would contain enough examples of the lexical item to be practised. Given the presence of all these factors, she was able to construct gap fill activities, but she found it impossible to ensure automatically that the content was appropriate to the purpose, so if this method of constructing language learning activities was followed, manual checking of the material would be necessary.

Therefore, although automatic generation of language learning exercises based on language use in corpora has the potential reduce the effort required in the construction of such exercises, it is still time-consuming, especially since the problems that Wilson found, such as the need for large tagged, graded and lemmatised corpora, all require extra work by the teacher if they are missing. However, if the exercises are not class-specific, they can be reused.

6.1.2 Vocabulary Features

The application of findings regarding vocabulary can be utilized to form a lexical syllabus (Flowerdew, 2012, pp. 194-5) based around lexical patterns as a component of teaching, and the content of this syllabus can be informed by indirect applications of corpus linguistics. One possible example of such an indirect application of corpus findings is the use of keyword analysis, based on a comparison of a reference corpus and a corpus of authentic instances of the target language. In the preceding chapters it has been pointed out that this can be of general academic writing and/or genre-specific academic writing. For the latter, a number of sources of authentic texts are available. Both BAWE and CJA14 have details of the discipline that the texts are from, and these texts can be extracted and grouped into sub-corpora, which can then be analysed. Another source of discipline-specific academic writing is to use the software 'AntCorGen' (Anthony, 2018), which is a corpus generation tool that downloads academic journal articles from PLOS ONE research. AntCorGen can also analyse parts of speech and cluster similar sentences into subgroups to show patterns of language use.

Academics can thus generate discipline-specific corpora for their own use in syllabus and materials design, or for students to use in direct applications. However, the creation of a corpus not only needs to follow Sinclair's (2004a) design principles as described in Chapter 3.2.1, it also involves processing textual source documents into suitable formats. Potential corpus texts may be in a variety of formats, such as word-processed documents, internet pages, Adobe Acrobat pdf format, or plain text. The texts need to be described and indexed (Cheng, 2012, p.31), as the BAWE documentation does, and given systematic file names. The format of the texts needs to be consistent, and common formats are .txt files which are suitable for plain text documents, and those with markup schemes that are designed to work in plain text, such as part-of-speech tagging. Corpus files that contain non-ASCII characters such as Chinese symbols may require a Unicode version, as BAWE has. More complicated markup schemes can use the XML markup system, for example to mark sections such as appendices, as is done in BAWE. Utilities exist to aid in this process, for example AntFileConverter (Anthony, 2017) can convert PDF and Word (DOCX) files into plain text. Wordsmith 7 (Scott, 2017) has a utility called 'Text Converter' which can copy files from Microsoft Word and Excel, pdf, and rich text format into Unicode and text formats.

At the level of the text in the documents, decisions about content need to be made, such as whether to follow the example of PLEC in removing quotes from a leaner corpus because they are not the learner's own words, whether to remove references, of whether to have multiple versions of the texts with different features. The BAWE corpus has been normalised to replace smart quotes with single straight quotes because they are functionally equivalent and doing so makes searches easier. Wordsmith 7's 'Text Converter' can also handle curly quotes.

This ability to create discipline-specific corpora is useful for English teachers, who, according to Leedham (2014) 'are seldom from the same disciplinary background as their students, and may not always be aware of the wide range of responses possible within the disciplines' (p. 131), and, as Römer (2011) notes, 'may not be experts in the specific discourse they teach' (p. 209) if they are teaching English for Specific Academic Purposes (ESAP) and are handling the writing of a discipline that they are not a member of. In order for students to be taught the appropriate language use for that discipline, as Hyland (2008b) emphasises, Leedham's (2014) suggestions could be followed, which are that ‘Investigation of disciplinary writing could take the form of corpus study of varying degrees of complexity, partnership activities between writing tutors and discipline lecturers, or could be as straightforward as reading assignments from their students’ disciplines’ (p. 129).

In a corpus study such as that which Leedham suggests, it would be possible to generate discipline-specific keyword lists through the comparison of an expert corpus, for example built using ‘AntCorGen’ (Anthony 2018), with a corpus of learner writing. However, caution would be necessary when creating such keyword lists, as it is important to retain details of not only the words, but the frequency with which they are used, their connotations, colligations, and where they fit into the move structure of the text. Such details are important for lexical priming (Hoey, 2005) so that students become familiar with the phraseology of academic texts, and this can be included in teaching either by extensive reading, or by direct instruction with multiple examples of the phrases in a more narrow context, which would be better in providing the amount of exposure to a phrase necessary to cement such priming.

These factors such as frequency can then be used to judge the teach-ability and teach-worthiness of the items, as described above for the Academic Formulas list (Simpson-Vlach and Ellis, 2010). If the frequency of use of a lexical item falls too low, as for example did that of *take / be taken into account*, which was used about once every twelve essays in BAWE-EON, the priority of teaching it can be reduced. However, frequency of use does not equate directly with usefulness, and if a lexical item uniquely or most frequently fulfils a function, its priority can be raised.

Another factor that concerns whether a lexical item is teach-worthy is that the same word may have different meanings in general English and in discipline-specific English. Mudraya (2006) gives the examples of *current, solution* and *tension*, which have specific meanings related to electronics, chemistry and engineering (pp. 239-42) and different collocates, for example, *alternating current, a salt solution,* and *constant tension*. Therefore the teaching of these and similar items to students from these disciplines could differ from teaching to a generic EAP class.

The distinction between receptive and productive vocabulary should also factor into the prioritisation of vocabulary items for teaching, both because there are less productive items, but also because they take longer to learn. If a pedagogic corpus is built and analysed for key words, those needed for mainly for productive use should be able to be separately identified from those needed mainly for receptive use.

Another corpus study that could contribute to the development of a lexical syllabus (Flowerdew, 2012) might involve the analysis of a pedagogical corpus for n-grams. Teachers could investigate the commonest n-grams in corpora of writing relevant to their students, either for EGAP or ESAP, and then provide lists of the more useful and teachable ones to their students. Due to the corpus-driven method of finding computationally-discovered n-grams, some of them are not ‘lexically whole’ (Leedham, 2014, p. 12), such as *that there is a* and *be seen in the* from Chen and Baker’s (2010) list, making them less teachable. Others would require careful teaching, for example *of a number of* and *in the number of* from Chen and Baker’s FLOB-J list should both be followed by plural nouns, but my students sometimes think that the following noun should agree in number with *a number*, and follow it with a singular.

6.1.3 Content

Teachers and subject leaders can investigate content concerns in a number of ways, including analysis of pedagogical corpora, and investigation of recurrent word combinations for EGAP contexts, and genre analysis of the phraseological profiles of discipline-specific corpora for ESAP contexts.

It was found from the phraseological profile of the PLEC students' writing that they were over-using language from the essay prompts, probably due to their restricted vocabulary. This problem could be addressed by teaching strategies for synonymy, such as referring to a subject by its hypernym, or name for the larger group to which something belongs, for example, a *questionnaire* is a type of *survey*. Because hypernyms can refer to multiple things in their group, they have wider use than synonyms. This is better than teaching direct synonyms, which can be problematic, because synonymy depends not only on meaning, but also collocation, for example, *answer* and *reply* are synonyms, but on a test students should *answer* the questions, not *reply* to them.

Disciplinary variation was found in the essay genre in BAWE by Nesi and Gardner (2012, pp. 94-130), and in the analysis in this thesis of the BAWE-EON and PLEC corpora. Indirect applications of corpus linguistics to disciplinary variation could include the analysis of corpora of discipline-specific texts such as journal articles and textbooks, using techniques such as keyword analysis to identify the content language that students may need for their studies. Comparison of expert with learner corpora can identify areas in which writers can develop.

6.1.4 Organisation

In the same way that computer programs such as word processors currently check spelling and grammar, other software can assist students to check their work for organisational issues, and provide information about lexical choices such as connector use. Tools such as this author's Common Error Detector webpage (http://www2.elc.polyu.edu.hk/cill/errordetector.htm) can be programmed detect problems that corpus analysis highlights, and warn students. For example, the Common Error Detector warns users about style if they use the informal organisational connectors *besides* or *last but not least*.

6.1.5 Conventions

The issue of whether rhetorical questions are appropriate in academic writing is a controversial one, with Milton (2001) being against it, but about 20 examples being found in the BAWE-EON corpus. The appropriacy of this may be discipline-dependent, or may even hinge on the attitude of the individual reader. To find out if it is discipline-dependent, teachers can search a discipline-specific corpus using the following regular expression: [A-Z][^\.\!\?]+?\byou\b[^\.\!]+?\? , which looks for a capital letter, probably at the start of a sentence, then a number of characters that are not full-stops, exclamation or question marks, followed by *you*, then some more characters, until a question mark is reached. The results of this search then need to be manually scanned to filter out non-rhetorical questions, such as in interview transcripts. From the remaining concordance lines, if any, the appropriacy of the questions can be judged, and a decision made as to whether this is worthy of teaching. Inter-faculty liaison may be necessary if the appropriacy could be reader-dependent, some academics may accept rhetorical questions that express doubt, some may accept them as a personalisation and persuasive strategy, but others may reject them as inappropriate to the genre or discipline. The educational background of the reader may also be a factor, as examples were found in American university argumentative essays, but not in British ones.

The Multidimensional Analysis Tagger (Nini, 2015) analysis based on Biber's (1998) dimensions does reveal that some students could incorporate more features of context-independent academic prose such as nominalisations in order to move their writing away from narrative register and towards scientific or learned exposition. The MAT tagger version 1.3 contains a tool to inspect a text for Dimensions features, which colour-codes a text according to Biber's dimensions, and tags it according to features that are factors in the dimension analysis, such as nominalizations and agentless passives. As the tool uses a considerable amount of grammatical and lexical meta-language, it is probably best suited to indirect applications, and the results of such investigations, for example a lack of nominalisation, pied-piping relative clauses or passive use, can be used to inform teaching.

Stance and voice are expressed differently across disciplines, according to Jiang's (2017) research. For example, he found that for "noun + *that*" structures in the category of cognition, such as *the view/idea/belief that*, we much more common in humanities subjects than in technical engineering and natural sciences, with figures of 8, 2 and 2 instances per hundred thousand words respectively. Therefore the indirect application of corpora is suggested in that teachers can investigate the discipline of their students and search for the noun structures given above and in Jiang's paper. A second reason that this might best be done by a teacher is the need to use a tagged corpus in order to ensure that the word matches are nouns, not verbs such as “I view that as a bonus”. A tagged corpus should be searched with a concordancer that can handle wildcards using search strings such as ‘view_NN that_*’ , because *that* can be tagged in different ways, including ‘IN’ for preposition or subordinating conjunction, or more confusingly ‘WDT’ standing for wh-determiner. These complications may prove overly challenging for learners or time-consuming to teach in class.

6.2 Direct Applications

Data-driven learning involves students using corpora for their own learning, and taking on the role of 'linguistic researcher' and 'language detective' (Johns, 2002 and 1997 respectively). This can be teacher-led, in which the teacher identifies features for investigation and instructs students on how to analyse the feature, or, after students have gained proficiency in such analysis, it can be a form of autonomous learning, such as when students search corpora to discover how to use a word or phrase by analysis of samples in a corpus, usually through the use of a concordancer.

However, this student role of linguistic researcher challenges students' language skills, powers of observation and inductive reasoning (Flowerdew, 2012, p. 197). Therefore there is a need for student training in corpus consultation (Flowerdew, 2012, p. 219). Such training should result in progression through stages of corpus competence, as suggested by Charles (2011, p. 40). The first stage is 'corpus awareness' in which students know what a corpus is, what kind of information it can provide, and how to access free corpora. The second stage is 'corpus literacy' in which students can conduct corpus searches and interpret concordance data, and so answer their own queries. The final stage is 'corpus proficiency', in which students can build their own corpora, formulate advanced searches, and interpret complex results. This autonomy allows students to work on their own queries (Flowerdew, 2012, p. 208), and may be useful if students are expected to write in a variety of disciplines (Leedham, 2014, p. 36), as they can search for how language is used in each discipline.

The needs of students are addressed by Römer (2011, p. 215), who organises her comments around their willingness and ability to deal with computer corpora.

Willingness involves students' motivation, and Römer suggests materials that are tailored to their needs and field of study. In my opinion, there needs to be a balance between discipline-specific and general academic input for two reasons, firstly because a significant proportion of graduates do not go on to work in the fields that they have studied, and secondly because discipline-specific language can be seen as partly a sub-set of general academic language.

A second factor affecting student motivation in using corpus software such as concordancers is that investigation of language use is slow for beginners, and may result in misconceptions. Wu (2010) states that 'Students may reach a wrong conclusion about some grammatical features, or their interpretation of these rules is either too broad or too narrow' (p. 75). As Sinclair (2004b) points out, students can easily 'derive nonsensical conclusions from the evidence' (p. 2). To address this issue, the teacher can plan how to scaffold the students' attempts at pattern finding, for example by providing hints (Bennett, 2010, p. 61; Johns 1994, p. 301, Flowerdew, 2012, p. 198).

A third factor affecting student motivation may be learning style, as Wu comments that 'Some students just do not like this kind of learning style and some kinds of language items are better 'given' than 'discovered'. Personal learning preferences will definitely influence one's learning results' (p. 75). Learning style was also addressed by Flowerdew (2012), who stated that 'field-dependent students who thrive in cooperative, interactive settings may benefit from corpus-based pedagogy, whereas field-independent learners may not take to this inductive approach to grammar, preferring instruction emphasizing rules' (p. 220). The need for further research into the effects of students' aptitude, intelligence, motivation and cognitive style on the success of DDL is also suggested by Johns (1994, p. 312).

An important factor in motivating the students is the approach used by their teacher. Yoon and Hirvela (2004) comment on how the instructor in their study:

> presented corpus searches… as a problem-solving approach to the language side of L2 writing. To make the searches and subsequent class discussions more meaningful and more practical in value (relative to the students' language learning needs), he emphasized the use of content words, e.g. the reporting verbs, as well as the often troublesome grammatical features, such as prepositions, and demonstrated how the corpus could assist in such use. (p. 265)

One of the students in their study reported that 'Now even when I am writing e-mail, I open corpora. And you know, some phrases, sentences if you are not very confident [whether] it's right or not, then I check corpora' (p. 275). He said that he using the corpus 'increased my confidence in English writing because I know many people write in this way, so I have confidence if I follow it' (p. 276).

The ability of students to deal with computer corpora depends on a number of factors. Römer has concerns about the suitability of concordance analysis for beginning or intermediate learners with a limited vocabulary. The level of learners is not only related to their general proficiency, but also their familiarity with the genre and lexis of the text, especially those of academic journal articles in an unfamiliar discipline. The solution that Römer suggests is to use the equivalent of graded readers. A number of data sources are suggested by Leedham (2014, p. 129), such as creating a corpus of good student texts from students' work over the years, either written by students on the same course, or taken from sources of discipline-specific texts like BAWE or the Michigan Corpus of Upper-Level Student Papers (MICUSP), or by creating a corpus composed of texts that students will encounter in their studies and that have been composed with their reading level in mind, such as course notes. Reading from such texts would also be a way 'to mimic the effects of natural contextual learning' (Cobb, 1997, p. 314) in that language features would be shown in the context of genre and discipline. In addition, Boulton and Cobb (2017, p. 385) found evidence that tailor-made corpora were more effective that large public corpora, and Flowerdew (2012, p. 221) suggests simplification of the downloaded corpus texts for students of lower ability.

Such level-graded corpora can be presented in two ways, either on with a print-out of concordance lines on paper or by using a concordancer program on computer. The paper method has the advantages that firstly the teacher can select specific concordance lines that contain the language features necessary for the students' analysis (Bennett, 2010, p. 21), and secondly that there are no worries about computer and software availability, no need to set up computers with concordancers and corpora, and no need or valuable class time absorbed to teach the students how to use the programme. Römer comments that such DDL worksheets could then be distributed to other teachers via the internet in order to popularise DDL (2011, pp. 214-5).

A number of examples of activity types that incorporate concordance lines is given by Johns (1994, pp. 300-310). These include gap-fills for which the answers can be found in a list of concordance lines, gap-fills in sentences taken from concordance lines and in which a range of possible answers are given, gap-fills in which groups of sentences from concordance lines all have the same answer which reveals contrasts in the use of the words in the answers, and matching exercises in which learners need to match two halves of a sentence based on the context revealed by the vocabulary in the sentence which illustrates the use of the word positioned just before the end of the first half of the sentence.

The second presentation method, that of students using corpus linguistics software in class, has a number of advantages and disadvantages. The most effective approach to DDL, according to a meta-analysis of 64 studies in the corpus linguistics literature by Boulton and Cobb (2017, p. 385), seems to be using a concordancer hands-on, rather than through printed materials. The advantages of using a concordancer in class include the ability to search for any word, the ability to click on the node word in the Key Word In Context (KWIC) display of search results and get a fuller impression of the context than the single line of the KWIC display, and for some concordancers, the ability to see tags such as part-of-speech tags, that make it easier to identify colligation patterns. In addition, teaching students how to use a concordancer can empower learners to find things out for themselves and so help them become more independent learners (Römer 2011, p. 213).

There are three main disadvantages of computers compared to paper. First are the problems mentioned above of getting operating concordancers and corpora into the hands of enough students in the class, although this does not mean all students, as it is probably best if students work on concordance investigation in small groups so that they can discuss the patterns that they find, and get support from group members if they cannot find such patterns of word use (Bennett, 2010, p. 21; Flowerdew, 2012, p. 219).

The second disadvantage is the need to use class time for the important task of teaching the students how to use the programme (Cobb, 1997, p. 302). For example, although Cobb's Compleat Lexical Tutor is a valuable resource, it is partly because there are many powerful and useful features that there are many choices that need to be explained to students on his concordancer page at https://www.lextutor.ca/conc/eng/, such as the choice of corpus, the options for associated words, and the controls for sorting the results.

The third disadvantage is that the teacher has less control over the lines displayed in the KWIC format readout, potentially causing problems such as unexpected grammar issues (Wu, 2010, p. 74), and too many lines might overwhelm students (Yoon and Hirvela, 2004, p. 261) and introduce too many repetitions of one pattern at the expense of too few of another, making it difficult for the students to distinguish between the signal and the noise. The final disadvantage is that the concordancer lines cannot automatically include gaps or be split into matching exercises to form exercises.

The need for easy-to-use software packages is emphasised by Römer (2011, p. 216), and by Sinclair (2003), who in his introduction to the procedural steps of corpus analysis, comments that ‘looking ahead, it is clear that more and more of this methodical work will be done eventually by computer’ (p. xvii). A number of features might be developed in order to make concordancers more user-friendly.

Automatic tagging of concordance results without a teacher or researcher having to tag their own corpora would save time for the teachers and thus encourage them to use more corpus-based materials in class. Such tagging could involve a range of tag types, such as part-of speech, sentence structure, positive and negative connotation, degree of formality, different meanings of polysemous words, and error identification.

Colour-coding of words according to their tag would probably make pattern-recognition easier and help with Sinclair's first and second procedural steps: *Initiate* and *Interpret*. The ability to sort the concordance lines according not only to the alphabetical order of words to the left and right of the node, but also by tag, would be helpful in pattern recognition.

Sentence structure tagging would be useful when looking for colligation patterns, for example in my experience my students often do not know the difference between how to use 'because' compare to 'due to', and a sentence-structure tagged concordance could show how 'because' is followed by a clause, whereas 'due to' is followed by a noun phrase.

Automatic pattern detection would make concordance analysis much easier by assisting with steps 3 and 4 of Sinclair's procedure: *Consolidate* and *Report*. In order to help students build corpora of academic journal articles for their own use, for example in helping to write theses and final year projects, it would be useful if concordancers could transparently handle pdf-formatted files and Microsoft Word files.

Some of this work has already been done or is in progress. Colour-coding of word classes in concordance lines was done by Sealey and Thompson (2007). On-the-fly tagging of parts-of-speech in web-based concordancers can be done using the JavaScript library 'Javascript Part of Speech Tagger' (jspos), available at https://code.google.com/archive/p/jspos/ . Online error identification is available in LanguageTool (Naber, 2003) and Grammark (Fullmer, 2018), which are both open source. Automatic pattern detection is done to a certain extent in Cobb's concordancer, in that it gives information on collocations beneath the KWIC display, for example, a search for *broad* gives potential collocation with *range, very, all, too* and *work*. Files that would normally be downloaded in pdf format can be converted to text format for corpus use, for example, AntCorGen (Anthony 2018) automatically converts the files that it downloads from the PLOS One open-access journals into text format. However, concordancers would be easier to use if all of these possibilities were standard features.

The following sections examine the findings from the previous two chapters in which the comparison suggested that the learners' writing was different from that in the expert corpus, and suggest ways of utilising them in teaching.

6.2.1 Grammatical Features

In general, language learning activities and materials can follow the Materials Analysis Checklist (Bennett, 2010 p. 30), shown in Appendix Three. In this checklist, grammar materials should contain explanations, either deduced inductively by students from concordancing, or given by teachers from their own investigation. An example worksheet that does both is given by Wu (2010), in which the students are given an example grammatical pattern V + to N + that-clause, and shown that *suggest* can follow this pattern in the sentence ‘I suggest to Miss Johnson that she sit down on the chair and wait’, but *advise* and *recommend* cannot. Students at Charles’ (2011) ‘corpus literacy’ stage could then be given other grammatical patterns, such as V + _ing and V + that-clause, and asked whether examples of sentences containing *suggest, advise* and *recommend* using these patterns can be found in the corpus.

Bennett’s checklist also recommends that materials should also provide grammar in context, for example in concordance lines, and for computer-based rather than paper-based activities, the context can be expanded by clicking on nodes in concordance lines. The checklist also covers writing materials, and suggests that they should develop students’ knowledge of rhetorical patterns such as patterns of usage in the concordance lines, engage students in the writing process, provide opportunities of writing for both fluency and accuracy, and connect reading and writing. More specific examples that include these factors are detailed below.

The findings from the investigation of parts of speech lead to advice that Hong Kong students use the expression *in my (personal) opinion* sparingly. Teachers and students at Charles’ (2011) ‘corpus literacy’ stage can search expert corpora for replacement realisations of the function of giving opinions, such as *seems* and *appears*, which appear almost 500 times in BAWE-EON. Over-use of comparative adjectives such as *more* and *less* can be countered by students searching expert corpora or model texts for more formal expressions such as *a greater number of* or *a smaller amount of*. The Academic Phrase Bank (Morley, 2014, pp. 44-5) has two pages of examples on comparisons.

Students' over-use of the simple present and present progressive, and under-use of the present perfect can be addressed using samples taken from expert corpora, and explanations of tense choices. Students could then be given samples of low-grade texts containing authentic tense errors taken from a pedagogical corpus, or students' own previous work or coursework from a previous semester, and be asked to correct the errors.

Regarding syntactic complexity, more attention could be paid to subordination in sentence structure. Students' awareness of the sentence length could be raised by a comparison of their own writing and that in an expert corpus or good models, and then they could work on use of subordinate clauses in their own writing. This might increase the syntactic complexity of their writing and hence increase the mean sentence length to a figure closer to the BAWE-EON mean number of words per sentence.

The tools recommended above for detection of grammatical errors, such as Language Tool (Naber, 2003), can be used in a number of ways. If students are doing in-class writing activities on computer, they can use the online versions of these tools to scan their text for errors and correct them before submitting their work to the teacher or showing it to class. In addition, individual students can use the tools to reflect on their own common errors, make a checklist of them, learn how to avoid them, and then use the checklist for their writing, not only for English classes, but for other university subjects that they are submitting texts for.

In an example of indirect application of these tools, if a teacher is building a learner corpus of their own students' work, texts can be scanned either manually or with these tools and tagged with the errors. This can enable the analysis of common errors, that can then be considered for inclusion and prioritised in teaching materials (Bennett 2010, pp. 77-80).

Regarding the over-use of *have* and *having*, students who are proficient in the use of concordancers can identify their own uses of these words along with the object or phenomena that they have, and then use a concordancer and an expert corpus to search for verbs that collocate with the object or phenomena. For example, the PLEC-EAP corpus contains "they cannot have a high income", and a search using the WebCorp internet concordancer shows that *earn* collocates with *income*, so the verb could be replaced with a more specific one, giving "they cannot earn a high income". WebCorp was used because formal collocations are often low-frequency combinations, for which smaller corpora may not return results, but it has the disadvantage of not being academic. If students have access to a large expert corpora, they might also find suitable collocations there. The CJA14 corpus has the collocation of *earn* and *income*.

Besides the use of concordancing, another corpus linguistics-based tool is the dictionary. Online dictionaries can now incorporate a number of interactive features. For example, if students are overusing particles, such as in *point out*, they can use sites such as OneLook Dictionary search at https://www.onelook.com to not only find a definition, but also find synonyms using the 'Words similar to' feature, example sentences from the 'Usage examples' link, and collocations using the 'Words that often appear near' function.

6.2.2 Vocabulary Features

The findings regarding word length suggested that native speaker writers use more short words such as articles, determiners and prepositions, and more words of six or more letters than the PLEC students. If learners have access to concordancing software with a word list function, these statistics can be found, for example for a corpus of discipline-specific writing. Learners can also be reminded to proof-read for missing articles, determiners and prepositions, and include nominalised words, which tend to be longer.

Type-token ratios were found to scale with proficiency for Chinese learners, which may indicate a narrower vocabulary. In the limited time available on university courses, vocabulary needs to be selected that students will find useful. Such vocabulary can be general vocabulary, as found in the Academic Word list and Academic Formulas list, or it can be discipline-specific and norm-specific, and researched from sources such as pedagogical corpora or discipline-specific sub-corpora of expert corpora, for example in CJA14.

Analysis of PLEC students' use of words from the Academic Word List shows that PLEC students are using slightly fewer academic words, especially from the higher numbered lists of rarer words in list five and higher. Therefore it can be concluded that teaching of the AWL should be more emphasised. This can be done firstly by students learning the words that they are unfamiliar with, and these unfamiliar words can be found using online resources such as the Randomized Checklist AWL Test at http://www.readingandwritingtools.com/cawlt/awlchecklisttest.html , and definitions of AWL words, their pronunciation, and online concordancer-connected usage examples of the AWL words can be learned at sites such as Academic Vocabulary at http://www2.elc.polyu.edu.hk/cill/eap/wordlists.htm , and practised at sites such as Vocabulary Exercises for the Academic Word List which can be accessed at http://www.englishvocabularyexercises.com/AWL/ . This knowledge can then be applied by students revising their previous work to include this vocabulary, then incorporating it into their future writing, both of which might involve concordancer use to investigate the usage, collocations and colligations of the vocabulary.

An interesting way for students to see how AWL words are used in their field is explained by Bennett (2010, p. 72), who describes how students can go to amazon.com, find a textbook from their field that has the 'Look Inside' feature, and then use the 'Search Inside this Book' feature to find instances of selected words, including from the AWL. The webpage will give a list of instances, and clicking them shows an image of the page from the book, so the context can be seen. For producing worksheets, the instances can be copied, and give concordance lines of about 70 characters in length, and screenshots of the book pages could be taken (within the limits of local copyright laws). The figure below shows an extract from a concordance for *analyze* built using this method from the textbook *Doing Corpus Linguistics* by Crawford and Csomay:

Figure 6.2.2.1: Concordance for 'analyze' from Amazon.com's Look Inside Feature

Page vi PART 3 Building Your Own Corpus, **Analyzing** Your Quantitati…
Page xi with information on how to build and **analyze** their own corpora (Part 3)
Page 4 part of linguistic study focuses on **analyzing** language and explaining w…
Page 6 characteristics. These texts are then **analyzed** collectively in or…
Page 8 characteristics: • it is empirical, **analyzing** the actual patterns of use in…
Page 29 already identified features could be **analyzed** and discuss how they cou…
Page 35 give you practice in searching and **analyzing** these units of language.
Page 42 search for keywords could be that you **analyze** a) what kinds of…
Page 43 wordandphrase.info. Click on 'Input/**Analyze** text…

Although the search term entered was *analyze,* the program has also selected other forms of the lemma, for example *analyzing* and *analyzed*. This technique also works for phrases. An example is the phrase *on the other hand* from the Academic

Formulas List (AFL). The extract in the figure below contains not only examples of the search term, but also related expressions, such as *other* and *handed me*.

Figure 6.2.2.2: Concordance for 'on the other hand' from Amazon.com's Look Inside Feature

Page 16 variation in different registers. On the one **hand**, they focus on v…
Page 27 seems to be more repetition than in the **other** text. The question is: Why…
Page 29 are typically in relationship with each **other** and do not occur in a vacu…
Page 30 announced she had a present for me and **handed** me a Marriott ca…
Page 51 … other hand, the three-word sequences of on the **other** and the other h…
Page 52 of the sentence position of '**on the other hand'** in spoken and writt…
Page 53 independent of the actual words. On the one **hand**, POS tags can help y…
Page 55 could just uncheck the boxes for all **other** possible POS categories (…
Page 70 our thermostats, recycle, etc. On the **other hand**, we have sent our …
Page 106 influences the results? Hypotheses, on the **other hand**, are basicall…
Page 108 very low that we are wrong. If, on the **other hand**, we must accept the …
Page 128 have two or more levels each. As with **other** parametric tests, the dep…

Therefore, this concordance may need to be filtered by the teacher before being used by students, or can be used without filtering to compare related language.

A more academic way for students to see how words are used in their discipline, and in a specific genre as well, is to use the Wordtree webpage (Stephenson, 2013) at http://wordtree.coventry.ac.uk/?BAWE . This page allows students to select BAWE texts from specific disciplines, genres and by the first language of writers, and will produce a tree diagram of the top 10 bi-grams of any word in the corpus. Each of these 10 tree 'branches' can be expanded to find the next top 10 trigrams, and so on. For example, a search for *that,* and then continuing by clicking on each top branch, gives the n-gram *that the majority of the arguments*. When there is only one example of the n-gram left in the corpus, the full sentence from the corpus is given; e.g. "that the majority of the arguments relating to interrogative clauses have already been outlined in section 2 the remainder of this essay will focus on the behaviour of quantifiers in negative clauses". Thus this tool, freely-available online, can help students to find the common phraseology of BAWE students' writing in the discipline and genre that they are working on.

Learners can also research statistics about their and others' texts using the free software for individual, non-profit research purposes called AntWordProfiler (Anthony, 2012), which can analyse types, tokens, and words in the AWL. Users can also view text from individual files, with the words colour coded according to vocabulary level. These can be edited using a thesaurus of synonyms, which may help with Staples' and Reppens' (2016) advice that learners need to find 'ways to avoid excessive use of the same noun phrases for cohesion' (p. 31).

There do not seem to be any online exercises for the Academic Formulas List (AFL). Texts, whether from a corpus or by learners, can be checked for phrases from the list by using a concordancer that can handle lists. For this research, Wordsmith 7 was used, as its Concord section has a function for searching a corpus for a list of phrases contained, one-per-line, in a text document. However, the resultant concordance list does not present the KWIC format readout in alphabetical order of the phrases, or in order of usefulness such as Simpson-Vlach and Ellis' Formula Teaching Worth. To do that, the KWIC file can be exported to a spreadsheet, and re-ordered there. However, as doing this in class would be complicated and time consuming, but mainly because most of the phrases were under-used or not used at all in the learner corpus, this stage could be omitted. The phrases could then be taught in order of utility, which the researchers have conveniently provided with their Formula Teaching Worth figures. Such teaching could include input on the meaning and use of the phrases, directed practice with gap fills made from the sentences in the KWIC display from a concordancer (Bennett, 2010, p. 67), followed by revisions to past texts and inclusion of the phrases in future ones.

Concerning students' repertoire of recurrent word combinations, Adel and Erman's (2012, p. 90) finding that native speakers tended to use more unattended 'this' constructions, hedges, and passive constructions was confirmed. Students can be taught to use *this* for anaphoric reference, taught to use sentence-initial *There* correctly in terms of agreement and sentence structure, taught hedging for academic style, and taught that passive voice is a common feature of academic writing, and how it should be used, for instance by examining a concordance read-out and analysing uses such as fronting of information in a sentence.

The n-grams found by Ebeling, such as *allows the reader to, as can be seen, at the beginning of, by the use of, can be seen in, could be argued that, it could be argued, the beginning of the, the importance of the, through the use of, to the fact that,* and *way in which the* (2011, p. 60), occurred much more rarely in PLEC. Such phrases could be added to the phraseology taught to students, using similar methods to those suggested for the AFL above.

A number of challenges to teaching phraseology are outlined by Byrd and Coxhead (2010), ranging from corpus construction to teaching and learning concerns. Firstly, in a similar analysis to that of single-word wordlists in Chapter 4, the composition of the corpora from which the phrase list is drawn is critical, including in terms of genre, level, representativeness and selection procedures. Secondly, some phrases occur inside others, for example as was mentioned above for *on the other hand*, in the Academic Phrase List *on the other, on the other hand,* and *on the other hand the* are counted as different sequences, and the teacher should decide which to teach. The third challenge is the lack of context for the phrases, for example, whether it is used at the start of a sentence, and what it collocates and colligates with. Byrd and Coxhead's final challenge is persuading students to read enough to encounter the phrases sufficiently frequently to acquire them, because learners need to read a phrase many times before it is incorporated into their lexicon.

To handle these challenges, it is important to select phrases drawn from a suitable corpus, to select phrases that are most teachable, to find samples with context, for instance by using a concordancer and clicking the node to see the wider context. To motivate students Byrd and Coxhead suggest the following strategies: selection of phrases which are immediately useful to students, exposure of students to multiple instances of the phrase over time, and implementation of general vocabulary learning strategies such as vocabulary notebooks and opportunities for the use of the phrases in both receptive and productive situations. For receptive skills they suggest the use of concordancing to collect sample texts that contain the phrases.

The findings of the analysis of the readability of the learners' texts showed that the sentence lengths were significantly shorter than those in the expert corpus. Students' attention can be drawn to the norm for academic writing of about 25 words per sentence, and two general approaches can be suggested: firstly, utilisation of more sophisticated sentence structures, for example using subordinate clauses, and secondly, using some lengthy academic phrases, such as some of those in Morley's (2014) University of Manchester Academic Phrasebank.

6.2.3 Content

Phraseological profiles of the corpora showed that grammatically-rich co-occurrences were most common, while lexically rich ones were less frequent. When analysing corpora, teachers should be aware of this limitation, and that the lexical co-occurrences are strongly influenced by topic, and therefore look for supporting confirmation from other sources of the usefulness of the phrases before considering them for inclusion in teaching materials.

Regarding the direct application of corpus techniques to disciplinary variation in content, students at Charles' (2011) 'corpus proficiency' stage can be shown how to create their own corpora specific to their needs, especially when their studies become more specialised in their discipline, for example in a thesis or final year project. In these contexts less teacher instruction and assistance may be possible, as each student or group will have an individual topic. Having a collection of models to refer to, for example the texts of the academic sources that are referred to in the literature review, and which can easily be searched for examples with a concordancer, can be of assistance to a more advanced learner writer, as Yoon and Hirvela's (2004) student explained above.

6.2.4 Organisation

Due to the fact that *therefore* and *however* are used much more in academic writing than in other genres (Conrad, 2000, p. 550), students can be encouraged to use them to a greater extent. Students at Charles' (2011) second stage of 'corpus literacy' could search expert corpora for examples of how and where to use them. Regarding the use of informal connectors, the replacement of *besides* and *last but not least* with more formal versions such as *in addition,* and *finally* could be added to students' personal proof-reading checklists.

6.2.5 Conventions

Referencing and citation skills are already a part of EAP subjects, and thus do not need to be reiterated here. However, the findings on dimensions of linguistic variation in register and genre are an issue.

The analysis of lexical items found in learners' academic writing which are more suitable for spoken discourse, based on Gilquin and Paquot's (2008) research, highlighted significant differences in the use of a number of lexical items. Suggestions for teaching materials could include more formal alternatives, for example, the use of *perhaps* instead of *maybe*, expressing possibility through the use of *apparently* and *presumably,* and the less frequent use or avoidance of a number of over-used informal, idiomatic or subjective expressions that are more common in spoken varieties of English, such as *Let's, every coin has two sides,* and *I think*. Gilquin and Paquot (2008) highlight the importance of wordlists that differentiate between synonymous lexical items in terms of the text types that they are appropriate for or whether their use is primarily spoken or written. To do this, the attention of students' at Charles' (2011) 'corpus awareness' stage can be drawn to examples from suitable corpora, after which students can edit their own or example texts, before going on to include under-used items of suitable register in their own writing, at a frequency matching target texts.

For students at Charles' (2011, p. 40) 'corpus proficiency' stage, analysis of stance and voice might be a feasible topic to introduce students to the possibility of tagging their own corpora, and searching them using regular expressions. Staples and Reppen (2016, p. 32) recommend discussion of ways to express stance with Chinese students. Students could be scaffolded with examples of "noun + *that*" constructions, before moving on to other part-of-speech tags, and possibly other tagging schemes such as those used by sentence parsers.

6.3 Summary

This chapter has discussed how the findings of this research could be used in practice, such as in decisions on course content and materials design. The issues were divided into indirect and direct applications of corpus linguistics, the former addressing pre-teaching issues, and the latter examining data-driven learning.

Selection criteria were discussed, including pedagogical relevance, and student and teacher needs. Suggestions of resources were given, including online concordancers, and other software tools. Examples of motivating resources, and how to create them, were given, including those appropriate for the students' level. Advantages and disadvantages of computer-based as compared with paper-based presentation were discussed. Ease of use of concordancers was examined, and suggestions were made for useful features in the software.

The findings from the previous chapters were then discussed, and suggestions made regarding strategies for including the implications into teaching.

Referring to the first research question of this thesis, the considerable extent of the differences between the BAWE-EON corpus and PLEC shows that changes to the content of EAP programs are advised, and this chapter has answered the second research question by suggesting ways in which this could be done. The significance of this was to provide a literature-based, pedagogically-appropriate series of suggestions for implementing the results of this corpus-based analysis into teaching and learning.

Chapter Seven: Conclusion

This chapter summarises the research, assesses its strengths and weaknesses, details its contribution to the literature, and suggests possible directions for further research.

7.1 Summary of the Research

The background to this research is that academic essay writing is an important skill for university students, and corpus linguistics contributes to the development of this skill by comparing expert and learner performance to identify areas for improvement.

The research gap is described by authors such as Evans and Morrison (2012) and Leedham (2014), who have found that content-area professors take little account of English skills, and writers such as Hyland (2015b) have found that feedback from faculty members on their students' writing rarely supports the development of writing in disciplinary-approved ways. Therefore it is up to EAP teachers to foster English skills, both for general EGAP and discipline-specific ESAP. The research gap is the one highlighted by Hyland (2008b), that further study in this area can help teachers and students to understand how writers can better employ the resources of English in different academic contexts. Filling this research gap can lead to improved academic writing by students, leading to greater satisfaction of students with their English abilities, the lack of which was pointed out by Evans and Morrison (2012). It would also help the students to fulfil their teachers' expectations in terms of academic writing.

Therefore this research has undertaken analysis of the academic essay genre, and discipline-specific analysis of the writing produced in various academic fields. This analysis has been implemented through the framework of Contrastive Interlanguage

Analysis (Granger, 1996) and the updated version called CIA^2 (Granger, 2015), which involved a corpus of advanced non-native English essay writing called PLEC, and comparison with a control corpus of comparable writing by native speakers of English, called BAWE. For comparability, a sub-corpus of BAWE was utilised that contained only the level one essays by native speakers of English, and this sub-corpus was called BAWE-EON.

From this research gap, framework and methods came the following research questions:

> To what extent are the commonly-taught aspects of academic essay writing and findings from the research literature on academic writing reflected in differences between the high standard essays by English native speakers in the BAWE corpus as compared to those of the Hong Kong students in the PLEC corpus?
>
> What changes would these differences (if any) suggest to the inclusion of these commonly-taught aspects?

This thesis is corpus-based in that it uses corpus linguistics 'to test existing theories or frameworks against evidence in the corpus' (Cheng, 2012, p. 6). The table below summarises the existing theories, the differences found between PLEC and BAWE-EON in the corpus comparison, and the suggestions for inclusion of new or revised teaching materials. It is ordered by the various aspects identified from the review of existing teaching materials. Further general suggestions are summarised below.

Table 7.1: Summary of the Research

Theories	Corpus Comparison	Suggestions
Grammar		
Parts of Speech		
In Hong Kong interlanguage, articles and prepositions are under-used (Milton, 2001)	Significantly under-used in PLEC	Teach phrases that include articles and prepositions, such as n-grams; e.g. *the importance of the*
Chinese speakers tend to use more pronouns than native speakers (Li, 2014)	Supported, *I* and *My* over-used in PLEC	Teach appropriate pronoun use for objectivity and stance
Collective pronouns indicating sense of belonging are over-used by Chinese students (Leedham, 2011)	Not supported, *We, us* and *ours* are under-used in PLEC.	Teach that collective pronouns should be used to identify with the academic and disciplinary community, not the wider public.
Phrasal verb / particle under-use by HK students (Milton, 2001)	Not supported, over-used in PLEC.	Students can use tools such as OneLook dictionary's synonym tool to find formal alternatives
Over-use of plural nouns by HK students, and plural nouns used only 1% less than singular and uncountable nouns (Milton, 2001)	All nouns over-used, but proportion of plural vs. non-plural similar to BAWE-EON	Do not discourage noun use because nominalisation is a feature of academic writing that PLEC students lack.
Comparative and superlative adjectives	Under-used in PLEC	Encourage use of more specific and multi-syllabic adjectives.
Over-use of *have* and *having*	Over-used in PLEC	Encourage use of more specific synonyms; e.g. *own, owning*
Tenses		
Chinese students under-use the present perfect (Hu and Gu, 2015)	Significantly supported	Highlight function-specific present perfect structures in corpora, such as for literature reviews: *X has been found to increase/ decrease with Y*
Syntactic Complexity		
Some measures of syntactic complexity may be reliably used to differentiate levels of L2 proficiency (Lu and Ai, 2015, pp. 16-7)	Supported for length of production unit, subordination, and clauses per sentence	Teach how to structure longer sentences, including subordinate clauses
Grammatical Errors		
Errors occur in BAWE as well as PLEC	Supported	Greater use of proof-reading tools
Fixed Multi-Word Constructions		
Constructions with fixed word order can be learned unanalysed, but constructions in which word order changes should be analysed.	Too few examples in the corpora to make it worthwhile	Not worth teaching unless it comes up as a problem in students' work

Theories	Corpus Comparison	Suggestions
Vocabulary		
Word Lengths		
Chinese does not have articles or phrasal verb particles, so the proportion of short words will be smaller.	Supported	See ‘articles and prepositions’ above.
Type-Token Ratios		
Type-token ratios indicates lexical density, and Chinese learners have less than native speakers (Staples and Reppen, 2016; Gui and Yang, 2001)	Supported. TTR also generally increases with proficiency.	Encourage vocabulary building that includes not only meaning, but how to use different forms of lexical items in sentences.
Norm-specific Vocabulary		
Legal norm-specific vocabulary is discipline specific (Breeze, 2011)	Supported for Breeze’s example words.	Teachers and/or students can create discipline-specific corpora and keyword lists.
Academic Word List		
Some vocabulary is more ‘academic’ and therefore could be expected to occur more frequently in higher-graded texts, i.e. BAWE	Supported for sub-lists 1, 3, and 5 – 10.	Students can do online tests of AWL knowledge, then search expert corpora to discover how to use items that they do not know, and then use them in their own writing.
Academic Formulas List		
Some phrases are more ‘academic’ and therefore could be expected to occur more frequently in higher-graded texts, i.e. BAWE	Supported, and academic phrase use scales with proficiency	
Repertoire of Recurrent Word Combinations		
Non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers. Adel and Erman (2012)	Supported for unattended ‘this’ constructions and passives. Not supported for existential ‘there’ constructions. Inconclusive for hedges.	Teach unattended ‘this’ constructions; e.g. for explaining reasons, and passives for objectivity.
N-grams		
Proficient language use involves the frequent use of phrases that can be found by n-gram analysis of native speakers’ texts. (Ebeling, 2011)	Supported, maybe because n-grams tend to include determiners and prepositions that are under-used in PLEC.	See ‘articles and prepositions’ above.
Use of formulaic expressions grows with writing proficiency. (Chen & Baker, 2010)	Supported at undergraduate level	See ‘articles and prepositions’ above.
Readability		
Readability scales with proficiency	Supported	Improve the students’ knowledge of long words and sentence structures including subordination.

Theories	Corpus Comparison	Suggestions
Content		
Phraseological Profiles		
Grammatically-rich co-occurrences at the top to the increasingly lexically rich as one moves down the list' (Cheng, 2012)	Supported. However, lexical n-grams are more common if the corpus' essays are on the same topic.	N-grams should be selected according to usefulness and teachability, not frequency.
Disciplinary Variation		
More use of personal pronouns, esp. *I* and *me*, in humanities such as philosophy, less in business and engineering (Nesi and Gardner, 2012)	Partially supported: true for business, not for engineering.	Use of personal pronouns depends on many factors, such as discipline, genre (e.g. argumentative essay) and reader expectation. Students should know how to decide.
Keyword analyses vary according to genre and discipline (Nesi and Gardner, 2012)	Supported.	Teachers and students could build their own genre- and discipline-specific corpus, and analyse it for keywords.

Theories	Corpus Comparison	Suggestions
Organisation		
Connectors		
Connector use is register-specific (Conrad, 2000)	Supported	A register-specific corpus can be scanned for informal connectors, such as *besides.* Students should know the register of various connectors.
Chinese students make greater use of particular connectors (Leedham, 2011)	Supported, some over-used, some under-used	In the teaching of connectors, not just their meaning, but also use, formality and frequency should be taught.

Theories	Corpus Comparison	Suggestions
Conventions		
Referencing and Citation		
Expert writers use more references and citations	Not supported	Not applicable.
Rhetorical Questions		
Over-used by Chinese students	Mixed. Varies over disciplines, language background and implicit answer to the question.	Students can be taught that the implicit answer to rhetorical questions in British English academic writing is one of doubt, according to Milton (2001), as opposed to their use in persuasive speeches.

Theories	Corpus Comparison	Suggestions
Register		
Over-use of spoken English features by learners, decreasing as proficiency increases (Gilquin and Paquot, 2008; Milton, 2001)	Generally supported, but with exceptions. Multiple factors that are difficult to distinguish.	Distinguish between spoken and written English. If in doubt, search a suitable corpus.
Dimensions of Linguistic Variation		
Corpora should conform to Biber's dimensional framework	Supported	Teachers can use the MAT software to analyse an expert corpora and discover the closest text type that students should aim for, and what features that that type has and the students' writing lacks.
Academic texts should become less general and narrative, and more scientific and learned, as proficiency increases.	Supported	Teach students to reduce the use of Dimension 2 contributing factors, including narrative features such as past tenses and third-person pronouns. They should aim for low scores in Dimension 1 and high scores in Dimensions 3 and 5. See Section 4.7.4 above.
Stance and Voice		
Expert essays will include more critical thinking (Durkin, 2011) operationalised through stance and voice features (Jiang, 2017)	Supported, with a few exceptional phrases	Students should adapt their expression of stance and especially voice to the cultural background of the reader, and use Jiang's phrases if appropriate.

Examining the patterns in the table above, it can be concluded that the results achieved from the analysis of one corpus or grouping of corpora may not be generalisable to other corpora, even if some features of the corpora are similar, for example, Milton's (2001) findings from his corpus of Use of English 'A' level writing are often different from an analysis of the PLEC corpus, despite the students having the same L1 background and therefore similar interlanguages. This is probably because only the highest-proficiency students from Milton's corpus would have gone on to university and become part of the academic stratum that wrote the PLEC essays.

Students' proficiency level and consequent progress in their interlanguage is a factor in many of the findings above, and language background seems also to be a factor, supporting Huat's (2012) analysis that 'some of the linguistic features of learner language are shared by learners from a wide range of L1 backgrounds while others are restricted to one particular learner population' (p. 193), and hinting that there are similarities in the interlanguages of non-native speakers from a variety of L1 backgrounds (Gilquin, Granger and Paquot, 2007), especially because learners are aiming for competence in a single underlying language, English, even if their aim is only to use it in an ELF context, or to use a non-native variety of English.

Disciplinary variation was shown in several of the findings, although general academic language uses also exist. Both seem important, as learners may need to operate in a number of disciplinary contexts and teachers may have mixed-discipline classes. It is suggested that teachers prepare curricula, lessons and materials specific to the students in their classes and their communicative needs, and advise them on what language is appropriate for which contexts.

Culture also seems to be a factor in some of the findings. Examples of this include that Chinese speakers tend to use more pronouns than native speakers (Li, 2014), seem more reluctant to express overt stance (Durkin, 2011), and use rhetorical questions somewhat differently (Milton, 2001). Writers can be advised to adapt their usage of such features to the culture of the reader that they are writing for.

Overall, as can be seen from Table 7.1 above, both similarities and significant differences were found between the literature on this subject and some of the findings. In addition, there were considerable differences between the performance of the PLEC and BAWE-EON students. Therefore, the use of corpus-based findings in materials design for teaching and learning is recommended, where such materials fit the needs of the students.

Chapter Six contained a discussion of the above findings, leading to recommendations for teaching and learning based on the literature. These included indirect applications, in which corpus linguistics was used outside the classroom, for example in planning curricula and materials, and also direct applications of corpus linguistics in data-driven learning by students using software such as concordancers inside the classroom.

The reasons for which this thesis does not include the creation of specific curricula, data-driven lesson plans, and materials are given. These include that subject content should be driven by the specific needs of a group of students, such a group not being available for the duration of this thesis, and resistance towards the use of corpora (Römer, 2011). Methods of overcoming this resistance were suggested, and convenient free resources and their functions were listed. Bennet's (2010) framework for creating corpus-designed indirect activities was explained, and examples given of the use of this framework and the resulting types of activity that teachers can prepare for students.

Direct applications of corpus linguistic techniques were also discussed, such as data-driven learning, in which students become linguistic researchers using tools such a concordancers. Charles' (2011) cline of corpus competence was used to stratify student activities, and practical methods to motivate students towards the use of these were given, such as the use of level-graded texts in corpora. Suggestions were made for the future of concordancers to enhance their ease-of-use. Specific examples of how the differences between the PLEC and BAWE-EON writers could be bridged were given in more detail than the table above, and related to principles in the literature, thus answering the research question ‘What changes would these differences (if any) suggest to the inclusion of these commonly-taught aspects?’

7.2 Strengths and Weaknesses of the Present Research

The strengths of the research include that it is based on quantitative evidence found in corpora, although interpretation of this evidence is qualitative (Leedham, 2014, p. 140). It does not depend on qualitative information from interviews and surveys regarding concepts that the interlocutors may well be unfamiliar with, as both students and content-subject teachers may not be familiar with the linguistic concepts under investigation. This does mean, however, that the author's interpretation is not reinforced or supplemented by additional evidence from interviewees.

The weaknesses of the research include weaknesses of corpus research in general and those specific to this research. The former includes the reliance on limited amounts of data, because 'assignments in the BAWE corpus were collected from just four universities (Oxford Brookes, Reading, Warwick and Coventry), and texts in each discipline were predominantly collected from a single university' (Leedham 2011, p. 266); and texts in the PLEC corpus were from a single university. This limits the generalisability of the research results.

When using regular expressions to automatically search corpora of learner texts, the number of matches is affected by the students' language errors, for example, spelling mistakes, which are likely to be more common in learner corpora . Search terms may not be matched for this reason. An example is searching for uses of words in the Academic Word List (Coxhead, 2000a) such as *accommodation* and its variants. The data showed that students used a variety of spellings with single and double 's's and 'm's. If the type of error is known in advance, the regular expression can be constructed to allow for it, but this is not always possible, for example due to the large number of words in the Academic Word List. Another problem is that spelling mistakes generate superfluous entries in word lists, making measures of keyness (Cheng 2012, p. 70) less reliable. They also affect statistics based on word frequency, such as type-token ratio and log-likelihood.

Spelling and grammatical errors also affect the accuracy of part-of-speech tagging. However, the extent of these errors was tested by identifying those in the F grade PLEC essays, and less than two percent of the words contained spelling errors. Only four tagging errors based on grammatical errors were found in the first 'F' grade essay. As shown by the high p levels in many of the findings in this research, the effects of these errors should be insignificant.

It is also unknown to what extent the BAWE students had received training in academic writing skills. The author's experience of British university education in the 1980s did not include such training, although the situation may well now have changed. Leedham (2014, pp. 122-4) surveyed 106 English university students in the UK, and found that although 90% said that they had received teaching on academic writing, many regarded this as vague or minimal, for example, only being given handouts. However, the PLEC students were all taking an academic writing subject. Therefore if the research is to be carried out on students from other language contexts, such aspects should be taken into account.

The thrust of much of the research into the BAWE corpus is genre-specific, such as Gardner and Nesi (2012) and Gardner (2008), but there is no parallel corpus of Hong Kong discipline-based high-scoring learner English of various genres, and it is beyond the scope of this thesis to compare and assess the relative efficiencies of generic and genre-specific approaches in course design. Such research is suggested in the directions for further research section below.

Much of the research in this thesis is quantitative, and the validity of the analysis could be reinforced by the use of a more qualitative approach, such as discourse analysis. However, McEnery, Xiao, and Tono (2006, p. 111) point out that there are major differences in text analysis amounting to a cultural divide between corpus linguistics and discourse analysis, although this is diminishing. In the interests of a manageable scope for this thesis, such qualitative analysis of the corpora has been left for further research.

7.3 Contribution to the Literature

This research could contribute to the literature by following up on the questions raised by Evans and Morrison (2012, p. 21) regarding the usefulness of university English courses. It could also help address the dissatisfaction among students with their English skills quoted in Evans and Morrison (2012, pp. 40-1).

The literature suggests a number of ways in which to address these issues based on a comparison of language between native and non-native English language learners. These include Liu's (2012, p. 33) fixed multi-word constructions; Nesi and Gardner (2006, p. 114) and Breeze's (2011) norm-specific vocabulary; Cheng, Greaves, Sinclair and Warren's (2008) examination of the phraseological profile of texts, Leedham's (2011) contrast of Chinese and English-speakers' academic writing, Adel and Erman's (2012, p. 90) repertoire of recurrent word combinations; Ebeling's (2011) n-grams; and Li's (2014) comparison of the use of first-person pronouns.

From these a number of recommendations can be made regarding the content of English courses. Based on the findings above, these recommendations include those in the table in the section 'Summary of the Research' above, such as more training in sophistication of sentence structures in order to increase the average sentence length; less use of pronouns, and especially *in my opinion*, advocating greater use of computerised proof-reading, and the replacement of informal expressions such as *besides, what's more,* and *last but not least* with more formal versions such as *in addition* and *finally.*

In addition to these recommendations on how to approach specific issues of word use, the thesis also contributes to a number of debates in corpus linguistics, for example, the generalisability of corpus comparison results (Leedham, 2011, p. 265), methods by which to judge students' proficiency level based on learner corpora (Lu and Ai, 2015, p. 16; Chen and Baker, 2010, p. 43), and the importance of disciplinary variation versus general academic language, also known as the 'common core hypothesis' (Flowerdew, 2012, p. 210; Coxhead, 2000b; Hyland and Tse, 2015).

Finally, it is hoped that this thesis can be of use to teachers and coordinators of university EAP subjects as a source of suggestions for how the methods and tools of corpus linguistics could be applied to the writing of their students, in an effort to identify areas for improvement in subject design, needs analysis, and materials design, and to provide suggestions for corpus-based student activities.

7.4 Directions for Further Research

Further research could be done in a number of areas, including theory, such as the ‘common core hypothesis’, and practice, for example new types of word list and new software capabilities.

The issue of the ‘common core hypothesis’, that there is a common core of academic vocabulary applicable to a range of disciplines, is a controversy that bears further investigation. For example, existing wordlists such as the AWL and AVL can be compared for common lexical items, which could then be applied to a range of corpora to investigate which gives the best coverage. This could be done for a range of corpora, genres and disciplines. Simpson-Vlach and Ellis’s (2010, pp. 493-6) Formula Teaching Worth could be applied to single-word lists, because both the lemmatised list approach and the word families approach have problems, as discussed above.

Regarding phraseology, a similar investigation could be done. For teaching purposes, the distinction between single-word lists such as the AWL and multi-word lexical items such as in the AFL probably matters little to students, and a combined list could be produced.

To take this lack of distinction concept further, if Sinclair (2003) is correct about there being little or no separation between grammar and vocabulary, and that these areas could be combined as lexicogrammar, it may be possible to create an academic lexicogrammar list. Certain grammatical structures are known to occur more frequently in academic prose, such as passive structures, as shown in Biber’s (1988)

Dimensions of Linguistic Variation in Register and Genre. Liu’s (2012) unfixed multi-word constructions also have a grammatical aspect, and would fit into such a list, as would Concgrams, in which components of the lexical items can be separated by intervening words and be in different order. The list would also incorporate the concept of colligation, as well as the collocation normally seen in sequential multi-word constructions.

One of the limitations of this thesis is that the PLEC corpus does not parallel the BAWE corpus in terms of genres, disciplines, topics and writers’ proficiency level. Further research could be done by the compilation of a Hong Kong equivalent to the BAWE corpus, maintaining the criteria that the texts should be high-scoring, so that a more direct comparison of academic writing styles and genres could be made.

Action research could be done with language learners on the user-friendliness of concordancers. For example, colour-coding of words according to their tag to make pattern-recognition easier and help with Sinclair's (2003) first and second procedural steps: *Initiate* and *Interpret,* could be the subject of experimentation to see if students could find patterns faster, saving valuable class time.

More student-centred corpus-linguistics based software could be developed. In my experience students often enquire about the difference between two words, and how to use them. Currently there is a webpage that can show two detailed dictionary definitions at http://www2.elc.polyu.edu.hk/cill/vocab/word_comparison.htm , but although the definitions are placed side by side, the comparison has to be done by the reader. Software could compare features of the two lexical items, including the meaning, collocation, colligation, and register, based on information in a large corpus, and then inform the user of the similarities and differences, with examples. To handle such a task, the corpus would have to have specialised selection criteria operating on a lexical-item basis rather than a text basis, to ensure that sufficient examples of the use of each lexical item in a variety of genres was available. It would also need to be tagged using multiple tagsets.

Finally, although the internet as a whole is not a corpus, due mainly to its lack of selection criteria for texts, it should be possible to build on the work of Anthony's (2018) AntCorpGen that generates corpora from online copyright-free texts in PLOS One journals, and build a tool to create corpora from a wider variety of copyright-free academic sources from the internet so that students and teachers can more easily create corpora personalised to their own needs. Such corpora could be assignment-specific, aiding the student with the necessary vocabulary, grammar, register, genre features, and disciplinary norms such as how to express stance and voice on the topic. This would help to address Hyland's (2015b) concern that although faculty staff would prefer their students to write in disciplinary-approved ways, writing instruction is often left to EAP writing teachers in EGAP classes, whose ability to offer assistance with assignments written for faculty staff is limited. It might also address some of the student concerns cited by Evans and Morrison (2012) regarding their writing style.

Appendix 1: L2 Syntactic Complexity Measures

These tables show the output of Lu and Ai’s (2015) ‘L2 Syntactic Complexity Analyzer’, which is a piece of software to measure language proficiency by assessing syntactical complexity. See Section 4.3.3 for details.

The first table shows the unnormalised counts of syntactic features. Scores for BAWE-EON are higher than PLEC because the former’s essays were longer. Comparing the PLEC scores at different grades, it can be seen that number of dependent clauses and complex T-units increases with grade.

		Syntactic structures																	
		Word count		Sentence		Verb phrase		Clause		T-Unit		Dependent clause		Complex T-unit		Coordinate phrase		Complex nominal	
Corpus		Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
BAWE-EON 150		1474	314	58	18	173	46	130	37	66	19	60	22	38	12	36	13	191	45
LOCNESS		824	469	Not given by Lu and Ai (2015)															
ICLE Chinese		515	107																
PLEC	A+	481	0	29	0	74	0	47	0	31	0	17	0	14	0	9	0	60	0
	A	539	65	28	5	71	10	50	7	29	5	20	5	14	4	10	4	73	13
	B+	486	150	27	10	63	21	46	15	28	10	16	7	12	5	11	6	63	22
	B	482	145	28	10	65	22	47	16	29	10	16	7	12	5	11	5	62	20
	C+	505	125	30	9	68	20	50	14	31	9	17	7	13	5	11	5	63	18
	C	461	157	28	10	63	23	45	16	29	11	15	7	11	5	10	5	56	21
	D+	443	137	27	10	61	21	44	16	28	10	14	7	11	5	10	5	54	19
	D	447	160	27	10	58	22	43	16	28	10	13	7	10	5	11	6	55	21
	F	428	93	30	2	57	4	42	3	30	4	12	2	8	1	10	6	61	14

It should be noted that because there is only one essay in the PLEC A+ grade, the standard deviations are zero.

The table below shows the length of production unit for various corpora. BAWE-EON 150 has the highest scores for mean length of clause, sentence and T-unit. This is followed by the texts in the LOCNESS corpus, while the data for ICLE Chinese and PLEC are similar.

The standard deviations for BAWE-EON 150 are also generally greater, showing more variability in length of clause, sentence and T-unit.

		Length of production unit					
		Mean length of clause		Mean length of sentence		Mean length of T-unit	
		Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
BAWE-EON 150		11.860	3.284	26.860	8.950	23.583	7.625
LOCNESS		10.092	1.624	19.602	4.321	17.308	3.210
ICLE Chinese		10.171	1.482	17.636	3.220	16.153	4.111
PLEC	A+	10.234	0.000	16.586	0.000	15.516	0.000
	A	10.886	1.435	19.903	2.958	18.663	2.562
	B+	10.764	1.543	18.509	4.038	17.572	3.663
	B	10.492	1.413	17.794	3.010	16.778	2.598
	C+	10.376	1.431	17.607	3.502	16.588	2.747
	C	10.403	1.900	17.159	4.936	16.259	4.859
	D+	10.615	2.793	17.807	8.738	16.631	5.617
	D	10.585	1.708	17.219	4.593	16.242	3.347
	F	10.135	1.742	14.198	3.305	14.628	3.387

The table below shows the amount of subordination in the corpora. The greater syntactic complexity of the BAWE-EON 150 essays is again shown, followed by that of the LOCNESS essays, and with ICLE Chinese and PLEC being similar.

		Amount of subordination							
		Clause per T-unit		Complex T-unit ratio		Dependent clause per clause		Dependent clause per T-unit	
		Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
BAWE-EON 150		2.002	0.360	0.586	0.124	0.453	0.085	0.935	0.337
LOCNESS		1.734	0.310	0.505	0.137	0.404	0.091	0.726	0.283
ICLE Chinese		1.600	0.227	0.436	0.127	0.345	0.078	0.567	0.192
PLEC	A+	1.516	0.000	0.452	0.000	0.362	0.000	0.548	0.000
	A	1.730	0.247	0.490	0.129	0.388	0.077	0.689	0.225
	B+	1.639	0.274	0.440	0.141	0.348	0.092	0.594	0.252
	B	1.614	0.250	0.424	0.135	0.341	0.092	0.571	0.232
	C+	1.612	0.253	0.420	0.127	0.336	0.084	0.560	0.222
	C	1.568	0.268	0.400	0.143	0.324	0.092	0.529	0.226
	D+	1.575	0.255	0.394	0.116	0.323	0.082	0.525	0.206
	D	1.546	0.275	0.361	0.146	0.305	0.093	0.493	0.221
	F	1.428	0.116	0.276	0.059	0.296	0.066	0.430	0.123

The following table on amount of coordination also shows a similar pattern.

		Amount of coordination					
		Coordinate phrase per clause		Coordinate phrase per T-unit		T-unit per sentence	
		Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
BAWE-EON 150		0.295	0.121	0.575	0.224	1.144	0.121
LOCNESS		0.253	0.113	0.429	0.175	1.132	0.121
ICLE Chinese		0.229	0.093	0.360	0.142	1.092	0.095
PLEC	A+	0.192	0.000	0.290	0.000	1.069	0.000
	A	0.207	0.086	0.344	0.124	1.066	0.058
	B+	0.235	0.100	0.379	0.157	1.055	0.076
	B	0.232	0.113	0.365	0.154	1.061	0.085
	C+	0.233	0.096	0.370	0.154	1.059	0.079
	C	0.236	0.100	0.363	0.146	1.056	0.104
	D+	0.231	0.119	0.355	0.171	1.059	0.111
	D	0.280	0.132	0.422	0.193	1.056	0.143
	F	0.246	0.129	0.365	0.202	0.974	0.059

The degree of phrasal sophistication shown in the following table shows a slightly different pattern, with the LOCNESS data showing similar figures to ICLE Chinese and PLEC. Complex nominals per T-unit seems to be a strong distinguishing feature of the BAWE-EON 150 essays.

		Degree of phrasal sophistication					
		Complex nominal per clause		Complex nominal per T-unit		Verb phrase per T-unit	
		Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
BAWE-EON 150		1.525	0.348	3.030	0.819	2.679	0.493
LOCNESS		1.222	0.330	2.087	0.565	2.342	0.434
ICLE Chinese		1.265	0.325	2.010	0.545	2.194	0.351
PLEC	A+	1.277	0.000	1.936	0.000	2.387	0.000
	A	1.493	0.337	2.556	0.567	2.438	0.328
	B+	1.394	0.297	2.281	0.613	2.252	0.438
	B	1.348	0.274	2.163	0.497	2.232	0.371
	C+	1.311	0.285	2.098	0.502	2.226	0.375
	C	1.277	0.309	1.987	0.551	2.206	0.443
	D+	1.289	0.332	2.008	0.509	2.220	0.413
	D	1.310	0.288	2.023	0.552	2.101	0.417
	F	1.433	0.245	2.062	0.458	1.941	0.187

This final table on overall sentence complexity shows the same general pattern of BAWE-EON 150 essays being most complex, followed by LOCNESS, with the ICLE Chinese and PLEC scores being similar at higher grades. PLEC scores show increasing sentence complexity with grade.

		Overall sentence complexity	
		Clause per sentence	
		Mean	St.Dev.
BAWE-EON 150		2.287	0.460571
LOCNESS		1.968	0.444
ICLE Chinese		1.748	0.301
PLEC	A+	1.621	0
	A	1.847	0.297491
	B+	1.730	0.326377
	B	1.711	0.288858
	C+	1.710	0.313049
	C	1.658	0.332595
	D+	1.673	0.355872
	D	1.637	0.386022
	F	1.387	0.083104

These figures seem to support the students' view given in Evans and Morrison's (2012) study, in which graduating students are cited as being 'far from satisfied with their English skills on graduation, lamenting… their unsophisticated writing style, limited repertoire of sentence patterns' (pp. 40-1). Since this seems to be a student concern, syntactic complexity should be considered for inclusion in academic writing instruction.

Appendix 2: A Comparison of Academic Formula List Sequences

The following table compares the top 181 Academic Formula List sequences taken from the sequences by Simpson-Vlach and Ellis (2010), ordered by degree of under-use in PLEC and then ordered by log-likelihood. Below these 181 the log-likelihood is below 3.84 and the significance is below 0.05. The first 48 sequences are not used in PLEC, so there are no log-likelihood, significance or p-levels for them.

It should be noted that different spellings of otherwise similar phrases are counted separately, and the combination of *appear to be* and *appears to be* would be the most frequent, with 239 occurrences. This use of *appear(s)* underlines the importance of hedging.

Of the sequences that do occur in PLEC, the top 3, starting at rank 49, are *can be seen, be seen as*, and *be argued that*, all of which are passive constructions, highlighting their under-use in PLEC. See Section 4.4.6 for details.

Rank	Word	BAWE		PLEC		Log-likelihood	Sig.	p-level	Log ratio	Use in PLEC
		Freq	Freq. / 100,000	Freq.	Freq / 100,000					
1.	in the form of	192	30.00	0	0					Under
2.	appears to be	153	23.91	0	0					Under
3.	needs to be	142	22.19	0	0					Under
4.	appear to be	86	13.44	0	0					Under
5.	we have seen	72	11.25	0	0					Under
6.	in the context of	72	11.25	0	0					Under
7.	in conjunction with	56	8.75	0	0					Under
8.	on the part of	55	8.59	0	0					Under
9.	insight into the	47	7.34	0	0					Under
10.	at the time of	41	6.41	0	0					Under
11.	as can be seen	35	5.47	0	0					Under
12.	has been used	35	5.47	0	0					Under
13.	in the absence of	34	5.31	0	0					Under

Rank	Word	BAWE		PLEC		Log-likelihood	Sig.	p-level	Log ratio	Use in PLEC
		Freq	Freq. / 100,000	Freq.	Freq / 100,000					
14.	the united kingdom	29	4.53	0	0					Under
15.	with respect to the	28	4.37	0	0					Under
16.	in this case the	28	4.37	0	0					Under
17.	same way as	28	4.37	0	0					Under
18.	by virtue of	28	4.37	0	0					Under
19.	take into account the	27	4.22	0	0					Under
20.	the same way as	27	4.22	0	0					Under
21.	to distinguish between	26	4.06	0	0					Under
22.	the presence of a	25	3.91	0	0					Under
23.	it follows that	23	3.59	0	0					Under
24.	assumed to be	23	3.59	0	0					Under
25.	in most cases	23	3.59	0	0					Under
26.	is determined by	23	3.59	0	0					Under
27.	degree to which	23	3.59	0	0					Under
28.	was based on	22	3.44	0	0					Under
29.	in more detail	19	2.97	0	0					Under
30.	shown in table	17	2.66	0	0					Under
31.	in both cases	17	2.66	0	0					Under
32.	similar to those	16	2.50	0	0					Under
33.	b and c	15	2.34	0	0					Under
34.	shown in figure	14	2.19	0	0					Under
35.	need not be	14	2.19	0	0					Under
36.	at the outset	12	1.87	0	0					Under
37.	none of these	12	1.87	0	0					Under
38.	in terms of a	12	1.87	0	0					Under
39.	in this paper	11	1.72	0	0					Under
40.	we assume that	10	1.56	0	0					Under
41.	the next section	10	1.56	0	0					Under
42.	is consistent with	9	1.41	0	0					Under
43.	can be expressed	7	1.09	0	0					Under
44.	in table 1	7	1.09	0	0					Under
45.	on the basis of the	6	0.94	0	0					Under
46.	be related to the	5	0.78	0	0					Under
47.	see for example	5	0.78	0	0					Under
48.	in the next section	4	0.62	0	0					Under
49.	can be seen	447	69.84	15	2.28	519.76	0.00	***	-4.94	Under
50.	be seen as	283	44.22	3	0.46	370.70	0.00	***	-6.60	Under
51.	be argued that	209	32.66	2	0.30	275.46	0.00	***	-6.75	Under
52.	as a whole	217	33.91	9	1.37	243.24	0.00	***	-4.63	Under
53.	it is important to	209	32.66	13	1.98	214.03	0.00	***	-4.05	Under
54.	an attempt to	151	23.59	2	0.30	194.79	0.00	***	-6.28	Under
55.	it is important	261	40.78	40	6.09	187.34	0.00	***	-2.74	Under

Rank	Word	BAWE		PLEC		Log-likelihood	Sig.	p-level	Log ratio	Use in PLEC
		Freq	Freq. / 100,000	Freq.	Freq / 100,000					
56.	it is possible	215	33.59	26	3.96	174.31	0.00	***	-3.09	Under
57.	it is clear	184	28.75	16	2.43	170.27	0.00	***	-3.56	Under
58.	can be seen in	116	18.12	1	0.15	153.77	0.00	***	-6.90	Under
59.	it has been	332	51.87	97	14.76	142.45	0.00	***	-1.81	Under
60.	it is clear that	140	21.87	9	1.37	142.12	0.00	***	-4.00	Under
61.	the nature of the	100	15.62	1	0.15	131.46	0.00	***	-6.68	Under
62.	be used to	155	24.22	17	2.59	131.21	0.00	***	-3.23	Under
63.	this does not	98	15.31	1	0.15	128.67	0.00	***	-6.65	Under
64.	it is possible to	131	20.47	12	1.83	119.01	0.00	***	-3.49	Under
65.	to the fact that	103	16.09	4	0.61	116.86	0.00	***	-4.73	Under
66.	it appears that	97	15.16	3	0.46	114.21	0.00	***	-5.05	Under
67.	to ensure that	91	14.22	4	0.61	100.87	0.00	***	-4.55	Under
68.	has also been	74	11.56	1	0.15	95.31	0.00	***	-6.25	Under
69.	can also be	149	23.28	29	4.41	91.76	0.00	***	-2.40	Under
70.	there has been	115	17.97	18	2.74	81.55	0.00	***	-2.71	Under
71.	it is interesting	59	9.22	1	0.15	74.57	0.00	***	-5.92	Under
72.	to do so	124	19.37	25	3.80	74.43	0.00	***	-2.35	Under
73.	with regard to	74	11.56	5	0.76	74.10	0.00	***	-3.93	Under
74.	take into account	60	9.37	2	0.30	69.84	0.00	***	-4.95	Under
75.	as a result of the	83	12.97	10	1.52	67.41	0.00	***	-3.09	Under
76.	depending on the	58	9.06	2	0.30	67.15	0.00	***	-4.90	Under
77.	which can be	109	17.03	23	3.50	63.20	0.00	***	-2.28	Under
78.	due to the fact	57	8.91	3	0.46	60.81	0.00	***	-4.29	Under
79.	are able to	114	17.81	27	4.11	60.09	0.00	***	-2.12	Under
80.	can be found	79	12.34	11	1.67	59.76	0.00	***	-2.88	Under
81.	as part of the	51	7.97	2	0.30	57.76	0.00	***	-4.71	Under
82.	due to the fact that	54	8.44	3	0.46	56.89	0.00	***	-4.21	Under
83.	high levels of	46	7.19	1	0.15	56.69	0.00	***	-5.56	Under
84.	factors such as	46	7.19	1	0.15	56.69	0.00	***	-5.56	Under
85.	can be used to	75	11.72	11	1.67	55.18	0.00	***	-2.81	Under
86.	be regarded as	47	7.34	2	0.30	52.43	0.00	***	-4.59	Under
87.	as a consequence	47	7.34	2	0.30	52.43	0.00	***	-4.59	Under
88.	it is interesting to	42	6.56	1	0.15	51.21	0.00	***	-5.43	Under
89.	for this reason	62	9.69	8	1.22	48.74	0.00	***	-2.99	Under
90.	this means that	82	12.81	17	2.59	48.19	0.00	***	-2.31	Under
91.	it is impossible to	50	7.81	4	0.61	47.58	0.00	***	-3.68	Under
92.	be explained by	43	6.72	2	0.30	47.12	0.00	***	-4.46	Under
93.	allows us to	41	6.41	2	0.30	44.48	0.00	***	-4.40	Under
94.	are based on	37	5.78	1	0.15	44.40	0.00	***	-5.25	Under
95.	little or no	36	5.62	1	0.15	43.04	0.00	***	-5.21	Under
96.	in accordance with	64	10.00	11	1.67	42.87	0.00	***	-2.58	Under
97.	an increase in the	43	6.72	3	0.46	42.67	0.00	***	-3.88	Under
98.	could be used	42	6.56	3	0.46	41.39	0.00	***	-3.85	Under

Rank	Word	BAWE		PLEC		Log-likelihood	Sig.	p-level	Log ratio	Use in PLEC
		Freq	Freq. / 100,000	Freq.	Freq / 100,000					
99.	it is possible that	47	7.34	5	0.76	40.30	0.00	***	-3.27	Under
100.	be noted that	37	5.78	2	0.30	39.23	0.00	***	-4.25	Under
101.	that there is no	69	10.78	15	2.28	39.08	0.00	***	-2.24	Under
102.	a wide range of	48	7.50	6	0.91	38.32	0.00	***	-3.04	Under
103.	a wide range	48	7.50	6	0.91	38.32	0.00	***	-3.04	Under
104.	his or her	39	6.09	3	0.46	37.58	0.00	***	-3.74	Under
105.	such as those	31	4.84	1	0.15	36.27	0.00	***	-4.99	Under
106.	it is impossible	61	9.53	13	1.98	35.09	0.00	***	-2.27	Under
107.	is based on the	42	6.56	5	0.76	34.30	0.00	***	-3.11	Under
108.	as shown in	33	5.16	2	0.30	34.02	0.00	***	-4.08	Under
109.	is likely to	87	13.59	28	4.26	33.36	0.00	***	-1.67	Under
110.	in a number of	32	5.00	2	0.30	32.73	0.00	***	-4.04	Under
111.	wide range of	52	8.12	10	1.52	32.30	0.00	***	-2.42	Under
112.	is more likely	34	5.31	3	0.46	31.30	0.00	***	-3.54	Under
113.	be considered as	34	5.31	3	0.46	31.30	0.00	***	-3.54	Under
114.	in this article	1	0.16	28	4.26	30.79	0.00	***	4.77	Over
115.	they did not	52	8.12	11	1.67	30.09	0.00	***	-2.28	Under
116.	at this stage	25	3.91	1	0.15	28.21	0.00	***	-4.68	Under
117.	been shown to	25	3.91	1	0.15	28.21	0.00	***	-4.68	Under
118.	that it is not	74	11.56	25	3.80	26.68	0.00	***	-1.60	Under
119.	should be noted	30	4.69	3	0.46	26.37	0.00	***	-3.36	Under
120.	it is not possible to	27	4.22	2	0.30	26.32	0.00	***	-3.79	Under
121.	is not possible to	27	4.22	2	0.30	26.32	0.00	***	-3.79	Under
122.	at least in	27	4.22	2	0.30	26.32	0.00	***	-3.79	Under
123.	two types of	32	5.00	4	0.61	25.55	0.00	***	-3.04	Under
124.	have shown that	29	4.53	3	0.46	25.15	0.00	***	-3.31	Under
125.	it should be noted	26	4.06	2	0.30	25.05	0.00	***	-3.74	Under
126.	was carried out	26	4.06	2	0.30	25.05	0.00	***	-3.74	Under
127.	it is not possible	31	4.84	4	0.61	24.37	0.00	***	-2.99	Under
128.	in accordance with the	22	3.44	1	0.15	24.22	0.00	***	-4.50	Under
129.	are likely to	63	9.84	21	3.19	23.11	0.00	***	-1.62	Under
130.	in such a way that	21	3.28	1	0.15	22.90	0.00	***	-4.43	Under
131.	such a way that	21	3.28	1	0.15	22.90	0.00	***	-4.43	Under
132.	does not appear	21	3.28	1	0.15	22.90	0.00	***	-4.43	Under
133.	the validity of the	21	3.28	1	0.15	22.90	0.00	***	-4.43	Under
134.	even though the	40	6.25	9	1.37	22.03	0.00	***	-2.19	Under
135.	the difference between the	23	3.59	2	0.30	21.28	0.00	***	-3.56	Under
136.	important role in	35	5.47	7	1.06	21.13	0.00	***	-2.36	Under
137.	can be found in	37	5.78	8	1.22	21.05	0.00	***	-2.25	Under
138.	does not have	65	10.16	24	3.65	20.73	0.00	***	-1.48	Under
139.	in some cases	60	9.37	21	3.19	20.64	0.00	***	-1.55	Under
140.	most likely to	19	2.97	1	0.15	20.27	0.00	***	-4.29	Under

Rank	Word	BAWE		PLEC		Log-likelihood	Sig.	p-level	Log ratio	Use in PLEC
		Freq	Freq. / 100,000	Freq.	Freq / 100,000					
141.	carried out by	49	7.66	15	2.28	19.95	0.00	***	-1.75	Under
142.	can be considered	24	3.75	3	0.46	19.16	0.00	***	-3.04	Under
143.	for the purposes of	21	3.28	2	0.30	18.81	0.00	***	-3.43	Under
144.	the purpose of this	21	3.28	2	0.30	18.81	0.00	***	-3.43	Under
145.	less likely to	32	5.00	78	11.87	18.63	0.00	***	1.25	Over
146.	their ability to	43	6.72	13	1.98	17.76	0.00	***	-1.76	Under
147.	small number of	20	3.12	2	0.30	17.58	0.00	***	-3.36	Under
148.	it is difficult	127	19.84	72	10.95	16.91	0.00	***	-0.86	Under
149.	to ensure that the	21	3.28	3	0.46	15.67	0.00	***	-2.85	Under
150.	in the course of	15	2.34	1	0.15	15.08	0.00	***	-3.95	Under
151.	there are a number of	38	5.94	12	1.83	14.91	0.00	***	-1.70	Under
152.	there are a number	38	5.94	12	1.83	14.91	0.00	***	-1.70	Under
153.	are a number of	38	5.94	12	1.83	14.91	0.00	***	-1.70	Under
154.	can be achieved	24	3.75	5	0.76	14.05	0.00	***	-2.30	Under
155.	a small number	16	2.50	2	0.30	12.77	0.00	***	-3.04	Under
156.	it is likely that	44	6.87	18	2.74	11.95	0.00	***	-1.33	Under
157.	there are no	59	9.22	29	4.41	11.25	0.00	***	-1.06	Under
158.	the most important	147	22.97	99	15.06	10.75	0.00	**	-0.61	Under
159.	is affected by	16	2.50	3	0.46	10.12	0.00	**	-2.45	Under
160.	a high degree	11	1.72	1	0.15	10.02	0.00	**	-3.50	Under
161.	it is necessary to	62	9.69	33	5.02	9.79	0.00	**	-0.95	Under
162.	give rise to	19	2.97	5	0.76	9.09	0.00	**	-1.96	Under
163.	is likely to be	32	5.00	13	1.98	8.79	0.00	**	-1.34	Under
164.	over a period of	10	1.56	1	0.15	8.79	0.00	**	-3.36	Under
165.	is that it is	35	5.47	15	2.28	8.77	0.00	**	-1.26	Under
166.	the other hand	245	38.28	320	48.68	8.08	0.00	**	0.35	Over
167.	we do not	48	7.50	26	3.96	7.24	0.01	**	-0.92	Under
168.	total number of	11	1.72	2	0.30	7.10	0.01	**	-2.50	Under
169.	the total number	11	1.72	2	0.30	7.10	0.01	**	-2.50	Under
170.	they do not	122	19.06	171	26.01	6.98	0.01	**	0.45	Over
171.	on the other hand	243	37.97	310	47.16	6.45	0.01	*	0.31	Over
172.	over a period	10	1.56	2	0.30	6.04	0.01	*	-2.36	Under
173.	be achieved by	10	1.56	2	0.30	6.04	0.01	*	-2.36	Under
174.	been carried out	12	1.87	3	0.46	6.03	0.01	*	-2.04	Under
175.	there are several	36	5.62	19	2.89	5.81	0.02	*	-0.96	Under
176.	it is necessary	78	12.19	53	8.06	5.49	0.02	*	-0.60	Under
177.	for this purpose	7	1.09	1	0.15	5.22	0.02	*	-2.85	Under
178.	can easily be	9	1.41	2	0.30	5.01	0.03	*	-2.21	Under
179.	it is worth	30	4.69	16	2.43	4.71	0.03	*	-0.95	Under
180.	on the other	290	45.31	353	53.70	4.61	0.03	*	0.25	Over
181.	be carried out	27	4.22	14	2.13	4.55	0.03	*	-0.99	Under

Note on p level: *** denotes p <.00001; ** denotes p <= .01; * denotes p < .05

Appendix 3: Materials Analysis Checklist

Bennett’s Materials Analysis Checklist (2010, p. 30) can be used as a guide for materials development for direct applications of the findings of corpus analysis. See Section 6.2.1 above for details.

Grammar Materials ➢ are logically sequenced ➢ exploit the three E’s (explanations, examples, exercises) ➢ provide grammar in context ➢ utilize both inductive and deductive reading	Reading Materials ➢ provide pre-, while-, and post-reading activities ➢ contain appropriate text types and topics ➢ use authentic texts, when possible
Speaking Materials ➢ consider the appropriate audience ➢ present grammar for the spoken context ➢ address accuracy and fluency ➢ address pronunciation ➢ provide speaking strategies ➢ link speaking and listening	Writing Materials ➢ develop students’ knowledge of rhetorical patterns ➢ engage students in the writing process ➢ provide opportunities for writing for both fluency and accuracy ➢ connect reading and writing
Listening Materials ➢ include strategies for listening ➢ allow for immediate post-listening production ➢ provide pre-, while-, and post-listening activities ➢ make use of appropriate spoken excerpts	

Appendix 4: List of Software Used

Anthony, L. (2012). *AntWordProfiler* (Version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2014). *AntConc* (Version 3.3.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2015). *TagAnt* (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2017). *AntFileConverter* (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2018). *AntCorGen* (Version 1.1.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Cobb, T. (2018). *Compleat Lexical Tutor v.8.3*. Retrieved from https://lextutor.ca

Free CLAWS web tagger (2017). Retrieved from Lancaster University, University Centre for Computer Corpus Research on Language Web site: http://ucrel-api.lancaster.ac.uk/claws/free.html

Frink, J. (2007). Flesch. [Computer software]. Retrieved from http://flesh.sourceforge.net/

Fullmer, M. (2018). Grammark. [Computer software]. Retrieved from https://github.com/markfullmer/grammark

Hüning, M. (2000). *TextSTAT version 2.9* [Computer software] Retrieved from http://neon.niederlandistik.fu-berlin.de/en/textstat/

Javascript Part of Speech (jspos) Tagger (2011). [Computer software] Retrieved from https://code.google.com/archive/p/jspos/

Ronald, S. (2015). *RocketReader Readability* [Computer software] Retrieved from https://sourceforge.net/projects/readability/

Scott, M. (2017). *WordSmith Tools version 7.0.0.127*. [Computer software] Stroud: Lexical Analysis Software.

Appendix 5: Sample PLEC Essays

A+ Grade Essay from the PLEC EAP corpus:

Importation of professionals from Mainland China into Hong Kong has raised concern of the public in general. Supporting views tend to say that the Admission of Mainland Professional Scheme would boost Hong Kong's competitiveness. On the other hand, voiced concerns were that the scheme may affect job opportunities of locals. To examine whether the scheme is beneficial to Hong Kong, advantages and disadvantages of importing professionals will be discussed.

Work opportunities could become limited with the importation of Mainland professionals. Unless more companies are set up, offering more positions, local graduates may have more difficulties in securing a job. It may be worse. Chan (1999) suggested that without a quota and a minimum wage, local university graduates would have a harder time seeking jobs. Employers, moreover, may be harsher on employees.

Feng (2001) argues that <q>. It may, therefore, be seen that job opportunities could be more instead of less with more talented personnel imported. Silicon Valley attracted people around the world and appeared to devolop faster with the group of talented people working together. Singapore goes further in welcoming professionals to work there bringing along their families.

Trade union leaders have contrasting views and urge for laws to be introduced protecting local workers. Wages in mainland China are likely lower than in Hong Kong. Thus, many appear to fear that they may be replaced by people from Mainland China who would be willing to do the same job at a lower pay. Cohen (2001, 21) pointed out that even professionals and managerial personnel needed protection from being unfairly laid off because of their higher wages.

Despite the fear that jobs might be taken away by mainland professionals, a survey has shown that companies have shortages and also have difficulties in filling the positions (Shamdasani, 2001). Importing professionals from Mainland China may reduce this shortage and allow time for local people to be trained with certain required skills. Some Hong Kong people are employed to work in Mainland China. Their jobs likely include training the workers in China. The same may be done in Hong Kong. Local workers may learn from imported professionals.

Moreover, it is common in universities to have teaching staff coming from various parts of the world. Such staff are also professionals, and students may learn from them. It may therefore be possible that talented people in other fields may teach people in those areas some new ideas or methods which may be refreshing.

Concluding what has been discussed, importing professionals has more benefits for Hong Kong Competition can lead to improvement, and thus secure or raise Hong Kong's status in the world. The government does, however, need to have a measure to make sure the wages are not lowered because of importing professionals. It also needs to confirm that such professionals imported are not available in Hong Kong.

F Grade Essay from the PLEC EAP corpus:

There are many people talking about "recycle" in the world nowaday, we can see the recycle signal on the televsion and poster. Even in some fast food shop. Their plastic cup is made by recycle material. What is the "recycle material"? It is a waste. For example, newspaper broken ploybag, broken appearl....etc. Those can recycle to product a useful thing again. There are so many countries to use the recycle policy, including hong kong. Is the recycling method of waste management have advantages in it. Please find out as below:

Recycling is a good method to reduce losing matedal in the world in hong kong. There are recycle policy is use the different colour box to collect three type of waste separatly. It is paper, tin, and plastic waste. After that, the separate waste will send to product the reuse produce. It can reduce waste management costs. As per wong & tang (2000. p.17 (2), 47 state that <q> in hong kong. There are large scale industrial invest in reuse produce since 1980's. However the recycling industrial have huge capital input . But lower income. Hughes (2000. p.71) suggest that huge capital input, high cost of collection and sorting, and inadequate facilities have hampered the development of the modern recycling industry.

Burning of plastic waste would made generates toxic and plastic waste is nonbiodergradable when landfilled.those method were only used to solve the plastc waste problem in hong kong before recycling. We can know those method was not a good method to used.

On the other hand the market demand for recycled products in hong kong is very small. Because most people like a popular thing. But the recycle products are not quite popular in here. No demand become no supply. It is no quite big commerical value therefore no bose want to develop this industry.

Recycling is a good method to save the world. But it must input huge capital but lower income. If one days, recycle products become a popular thing. Have a big demand. It will suitiable for hong kong.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. doi: 10.1016/j.jeap.2013.08.002

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes 31*, 81–92

Ahmad, K. (2005). 'Terminology in text', Tuscan Word Centre Workshop. Siena, Italy. June 2005.

Aston, G. (2009). The learner as corpus designer. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 7-25). Amsterdam: Rodopi.

Bennett, G. R. (2010). *Using corpora in the language learning classroom*. Ann Arbor: University of Michigan Press.

Bhatia, V. K. (1997). *Genre analysis today*. Revue belge de philologie et d'histoire, 75(3), 629-652.

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

Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1), 3-44. doi: 10.1515/ling.1989.27.1.3

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing, 8*(4), 243-257.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Bloch, J. (2010). A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers. *Journal of Writing Research*, 2(2), 219-244.

Bloor, M., & Bloor, T. (1986). Language for specific purposes: Practice and theory. In *CLCS Occasional Papers*. Dublin: Centre for Language & Communication Studies, Trinity College.

BNC2 POS-tagging manual: POS-tagging error rates (2015). Retrieved July 9, 2017 from University of Lancaster, University Centre for Computer Corpus Research on Language Web site: http://ucrel.lancs.ac.uk/bnc2/bnc2error.htm

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67* (2), 348–393.

Breeze, R. (2011). Disciplinary values in legal discourse: a corpus study. *Iberica 21*, 93-116.

Brown, R. (1973). *A first language: The early stages*. London: George Allen & Unwin.

Burstein, J., Chodorow, M. & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine, 25*(3), 27–36.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*(5), 31-64.

Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In A. Frankenburg-Garcia, L. Flowerdew & G. Aston (Eds.) *New trends in corpora and language learning* (pp. 26-43). New York: Continuum.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, (14) 2, 30-49

Cheng, W. (2007). Concgramming: A corpus-driven approach to learning the phraseology of discipline specific texts. CORELL: *Computer Resources for Language Learning* 1, 22-35.

Cheng, W. (2012). *Exploring corpus linguistics*. Abington, Oxfordshire: Routledge.

Cheng, W., Greaves, C., Sinclair, J., & Warren, M. (2008). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics 30,* 2: 236–252. doi:10.1093/applin/amn039

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System, 25*, 301–315.

Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185–206). Cambridge: Cambridge University Press.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*(3), 548-560.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal*, 52(1), 57-63. doi:10.1093/elt/52.1.57

Coventry University (2014). *British Academic Written English Corpus (BAWE).* Retrieved October 2, 2014, from Coventry University, Web site: http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/

Coventry University (2015). *Genres of academic writing in the BAWE corpus.* Retrieved June 25, 2015, from Coventry University, Web site: http://www.coventry.ac.uk/research-bank/research-archive/art-design/british-academic-written-english-corpus-bawe/contents-of-the-bawe-corpus/about-us/

Coventry University (2015). *The distribution of genre families.* Retrieved June 25, 2015, from Coventry University, Web site: http://www.coventry.ac.uk/Global/05%20Research%20section%20assets/Research/British%20Academic%20Written%20English%20Corpus%20(BAWE)/Images/Microsoft%20Wordwordlist%20explanation.docx%20wordlist%20explanation.pdf

Coxhead, A. (2000a). A new academic word list. *TESOL Quarterly*, Vol. 34, No. 2 (Summer, 2000), pp. 213-238. http://www.jstor.org/stable/3587951

Coxhead, A. (2000b). The academic word list: A corpus-based word list for academic purposes. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 72-89). Amsterdam: Rodopi.

Crawford, W., & Csomay, E. (2015). *Doing corpus linguistics*. New York: Routledge.

Crosthwaite, P. (2016). A longitudinal multidimensional analysis of EAP writing: Determining EAP course effectiveness. *Journal of English for Academic Purposes, 22*, 166-178.

Dejica-Cartis, D., & Cozma, M. (2013). Using Theme-Rheme analysis for improving coherence and cohesion in target-texts: a methodological approach. *Procedia - Social and Behavioral Sciences 84*, 890 – 894.

Dunleavy, P. (2003). *Authoring a PhD: How to plan, draft, write and finish a doctoral thesis or dissertation*. New York, NY.: Palgrave MacMillan.

Durkin, K. (2011). Adapting to western norms of critical argumentation. In L. Jin & M. Cortazzi (Eds.) *Researching Chinese learners* (pp. 274-291). Houndmills: Palgrave Macmillan.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes, 43*, 49-61. doi: 10.1016/j.esp.2016.01.004

Ebeling, S. O. (2011). Recurrent word-combinations in English student essays. *Nordic Journal of English Studies (NJES) 10* (1), 49-76

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3-17. doi: 10.1016/j.jeap.2006.11.005

Evans, S., & Morrison, B. (2012). Learning and using English at university: Lessons from a longitudinal study in Hong Kong. *The Journal of Asia TEFL*, *9*(2), 21-47.

Flowerdew, L. (2004). The argument for using English specialised corpora to understand academic and professional language. In U. Connor & T. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11-33). Amsterdam: John Benjamins.

Flowerdew, L. (2012). Corpora in the classroom: An applied linguistic perspective. In K. Hyland, M.H. Chau & M. Handford (Eds.), *Corpora in applied linguistics: current approaches and future directions* (pp. 208-224). London: Continuum.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics, 35*(3), 305-327. doi:10.1093/applin/amt015

Gardner, S. (2008). Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: an illustration through university history and engineering assignments. In Steiner, E., & Neumann, S. (Eds.) *Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop* 23rd - 25th July 2007, Saarbrücken, Germany.

Gardner, S. (2012). Perspectives on the disciplinary discourses of academic argument. In Groom, N. (Ed.), *Proceedings of International Corpus Linguistics Association Meeting 2011*: Discourse and Corpus Linguistics.

Gardner, S., & Nesi, H. (2012). A classification of genre families in university student writing. *Applied Linguistics 34* (1) 1-29.

Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, doi: 10.1093/applin/amy005

Gerritsen, M., & Nickerson, C. (2009). BELF: Business English as a lingua franca. In F. Bargiela-Chiappini (Ed.) *The Handbook of Business Discourse* (pp. 180-192). Edinburgh: Edinburgh University Press.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*(1), 41-61.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*, 319-35.

Granger, S. (1990). *UCL/CECL Centre for English Corpus Linguistics ICLE Corpus*. Retrieved from Université catholique de Louvain, The Louvain Centre for English Corpus Linguistics Web site: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm

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

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly 37*(3), 538-545.

Granger, S. (2015). Contrastive interlanguage analysis. A reappraisal. International *Journal of Learner Corpus Research 1* (1): 7-24. doi: 10.1075/ijlcr.1.1.01gra

Greaves, C. (2009). *ConcGram 1.0; A phraseological search engine*. Amsterdam: John Benjamins.

Gui, S. C., & Yang, H. Z. (2001). *Computer analysis of Chinese learner English*. Retrieved August 20, 2017 from Hong Kong University of Science and Technology, Center for Language Education Web site: http://cle.ust.hk/common/conf2001/keynote/subsect4/yang.pdf

Hamp-Lyons, L., & Heasley, B. (2006). *Study writing*. Cambridge: Cambridge University Press.

Hardie, A. (2014). *Statistical identification of keywords, lockwords and collocations as a twostep procedure*. ICAME 35: Corpus Linguistics, Context and Culture: Book of Abstracts, 49.

Hartley, J. (2008). *Academic writing and publishing: A practical handbook.* Abingdon: Routledge.

Heuboeck, A., Holmes, J., & Nesi, H. (2010). *The BAWE Corpus Manual*.

Hoey, M. (2005). *Lexical priming: A new theory of language*. London: Routledge.

Hu, S., & Gu, Y. (2015, November). A corpus-based study of present tense distribution in Chinese students' English writings. In *2015 International Conference on Social Science, Education Management and Sports Education*. Atlantis Press. doi: 10.2991/ssemse-15.2015.35

Huat, C. M. (2012). Learner corpora and second language acquisition. In K. Hyland, M.H. Chau & M. Handford (Eds.), *Corpora in applied linguistics: current approaches and future directions* (pp. 191-207). London: Continuum.

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

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, MI: University of Michigan Press.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes 27*, 4–21.

Hyland, K. (2012). Corpora and academic discourse. In K. Hyland, M.H. Chau & M. Handford (Eds.), *Corpora in applied linguistics: current approaches and future directions* (pp. 30-46). London: Continuum.

Hyland, K. (2015a). *Teaching and researching writing*. New York: Routledge.

Hyland, K. (2015b, January). *Feedback on writing: Faculty and student perceptions*. Symposium conducted at Hong Kong University.

Hyland, K., & Guinda, C. S. (Eds.), (2012). *Stance and voice in written academic genres*. Houndmills, UK: Palgrave Macmillan.

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

Hyland, K., & Tse, P. (2015). Is there an "academic vocabulary"? In H. Basturkmen (Ed.), *English for academic purposes: Critical concepts in linguistics* (pp. 235-253). Abingdon: Routledge.

Jiang, F. K. (2017). Stance and voice in academic writing. *International Journal of Corpus Linguistics, 22(*1), 85-106.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293-313). Cambridge: C.U.P.

Johns, T. F. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). London: Longman.

Johns, T. F. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 107–117). Amsterdam: Rodopi.

Jordan, R. R. (1997). *English for Academic Purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7-36.

Larsen-Freeman, D., & Long, M.H. (1991). A*n introduction to second language acquisition research*. New York: Routledge.

Learner Corpus Association (2014). *The Louvain Corpus of Native English Essays (LOCNESS) - Learner Corpus Association*. Retrieved July 26, 2016 from http://www.learnercorpusassociation.org/resources/tools/locness-corpus/

Lee, D. (2000). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 107–117). Amsterdam: Rodopi.

Lee, D., & Chen, S. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners, *Journal of second language writing 18,* 281–296. doi: 10.1016/j.jslw.2009.05.004

Leedham, M. (2011). *A corpus-driven study of features of Chinese students' undergraduate writing in UK universities*. PhD thesis The Open University (UK).

Leedham, M. (2014). *Chinese students' writing in English: Implications from a corpus-driven study*. Abingdon: Routledge.

Lexical Items (n.d.). Retrieved January 18, 2015 from http://www.coventry.ac.uk/research-bank/research-archive/art-design/british-academic-written-english-corpus-bawe/research-/lexical-items/

Li, D. C., & Luk, Z. P. S. (2017). *Chinese-English contrastive grammar: An introduction*. Hong Kong University Press.

Li, L. (2014). Contextual and cultural influence on the use of first person pronouns by Chinese learners of English. In Qian, D., & Li, L.(eds.) *Teaching and learning English in east Asian universities: Global visions and local practices* (pp. 302-322). Cambridge: Cambridge Scholars Publishing.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language learning, 54*(2), 193-226.

Lin, L. H. F. (2003). Corpus evidence: Existential constructions by Chinese learners of English. *The English teacher- An international journal*. *6* (3) 279-280.

Liu, D.L. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes* 31, 25–35.

Long, R. (2013). A review of ETS's Criterion online writing program for student compositions. *The Language Teacher, 37*(3), 11-18.

Louhiala-Salminen, L., & Kankaaranta, A. (2011). Professional communication in a global business context: The notion of global communicative competence. *IEEE transactions on professional communication*, *54*(3), 244-262.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics, 15*(4), 474-496.

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. doi:10.1016/j.jslw.2015.06.003

Malmström, H., Pecorari, D., & Shaw, P. (2018). Words for what? Contrasting university students' receptive and productive academic vocabulary needs. *English for Specific Purposes*, 50, 28-39. doi: 10.1016/j.esp.2017.11.002

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book.* Abingdon: Routledge.

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, *12*(8), 639-646.

McWhorter, K.T. (2010). *Pathways: Writing scenarios*. New York: Pearson.

Milton, J. (2001). Elements of a written interlanguage: a computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students. *Research Reports*, 2. Hong Kong: Language Centre, The Hong Kong University of Science and Technology.

Morley, J. (2014). *The Manchester academic phrase bank.* University of Manchester. Retrieved from: http://www.phrasebank.manchester.ac.uk/

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*(2), 235-256. doi:10.1016/j.esp.2005.05.002

Naber, D. (2003). *A rule-based style and grammar checker*. Diplomarbeit, Technische Fakultät, Universität Bielefeld.

Nesi, H., & Gardner, S. (2006). Variation in disciplinary culture: University tutors' views on assessed writing tasks. In: R. Kiely, G. Clibbon, P. Rea-Dickins & H. Woodfield (Eds.), *Language, culture and identity in applied linguistics* (pp. 99-117). (British Studies in Applied Linguistics, Volume 21) London: Equinox Publishing.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education.* Cambridge: Cambridge University Press.

Nesi, H., & Gardner, S. (2017). Stance in the BAWE Corpus: New revelations from Multidimensional Analysis. *Proceedings 9th International Corpus Linguistics Conference, Birmingham University* July 24-28 2017.

Nini, A. (2014). *Multidimensional Analysis Tagger 1.3 - Manual*. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

Nunan, D. (2003). The Impact of English as a Global Language on Educational Policies and Practices in the Asia-Pacific Region. *TESOL Quarterly, 37*(4), 589-613. doi:10.2307/3588214

Nunan, D. (2007). *What is this thing called language?* Basingstoke: Palgrave MacMillan.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oshima, A., & Hague, A. (2007). *Introduction to academic writing*. White Plains, N.Y.: Pearson.

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London & New York: Continuum.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149.

PLOS ONE: Accelerating the publication of peer-reviewed science (2018). *PLOS ONE*. Retrieved from http://journals.plos.org/plosone/

Renouf, A., Kehoe, A., & Banerjee, J. (2007). WebCorp: an integrated system for web text search. *Language and Computers, 59*, 47.

Research Centre for Professional Communication in English (2014). *Corpus of Journal Articles 2014*. Retrieved from The Hong Kong Polytechnic University, Department of English Web site: http://rcpce.engl.polyu.edu.hk/cja2014/default.htm

Research using the BAWE corpus (2018). Retrieved from Coventry University, Web site: https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/british-academic-written-english-corpus-bawe/

Römer, U. (2011). Corpus research applications in second language teaching. *Annual review of applied linguistics, 31,* 205-225.

Ronald, S. (2013). RocketReader Readability. [Computer software]. Retrieved from https://sourceforge.net/projects/readability/

Rosen, L.J. (2012). *The academic writer's handbook*. Boston: Pearson.

Salazar, D.J.L. (2008). *Lexical bundles in scientific English: A corpus-based study of non-native writing*. (Unpublished doctoral thesis). Universitat de Barcelona.

Sealey, A., & Thompson, P. (2007). Corpus, concordance, classification: young learners in the L1 classroom. *Language awareness*, *16*(3), 208-223.

Setter, J., Wong., C. S. P., & Chan, B. H. S. (2010). *Hong Kong English*. Edinburgh: Edinburgh University Press.

Seidlhofer, B. (2011). English as a lingua franca. *ELT Journal 59*(4), 339-341; doi:10.1093/elt/cci064

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics, 31* (4), 487–512.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2003). *Reading concordances: An introduction*. Harlow: Pearson.

Sinclair, J. M. (2004a). *Developing linguistic corpora: a guide to good practice*. Retrieved July 29, 2016 from King's College London, Arts and Humanities Data Service Web site: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

Sinclair, J. M. (2004b). *How to use corpora in language teaching*. Amsterdam: John Benjamins.

Skwire, D. (2012). *Student's book of college English: rhetoric, reader, research guide, and handbook*. New York: Pearson.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17-35.

Stephenson, P. (2013). *Wordtree* v.3.5. Retrieved from Coventry University, Web site: http://wordtree.coventry.ac.uk/?BAWE

Sublists of the Academic Word List (2010). Retrieved July 29, 2016 from Victoria University of Wellington, School of Linguistics and Applied Language

Studies Web site: http://www.victoria.ac.nz/lals/resources/academicwordlist/publications/awlsublists1.pdf

Swales, J. (2014). Variation in citational practice in a corpus of student biology papers: From parenthical plonking to intertextual storytelling. *Written Communication. 31*(1). 118-141.

Szudarski, P. (2018). *Corpus linguistics for vocabulary. A guide for research*. Abingdon, Oxon: Routledge.

The PolyU Language Bank -- General. (2015). Retrieved June 24, 2015 from The Hong Kong Polytechnic University, The English Department Web site: http://langbank.engl.polyu.edu.hk/general.html

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (pp. 173-180). Association for Computational Linguistics.

West, M. P. (Ed.). (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans, Green.

Widdowson, H. (2000). The limitations of linguistics applied. *Applied Linguistics 21*(1), 3-25.

Wilson, E. (1997). The automatic generation of CILL exercises from general corpora. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 116-130). Harlow: Addison Wesley Longman.

Wu, W. (2010). The integration of corpus-based data into grammar instruction: Using advise, recommend, and suggest as an example. *Journal of Education and Foreign Language and Literature, 8*, 67-84.

Xu, Jiajin. (2009). *Log-likelihood ratio calculator*. Beijing: National Research Centre for Foreign Language Education, Beijing Foreign Studies University.

Xue, G., & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication 3*, 215-229.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing, 13*, 257-283. doi: 10.1016/j.jslw.2004.06.002.