Investigating Lexical Bundles Across Learner Writing Development

YU-HUA CHEN

A Thesis Submitted to
The Department of Linguistics and English Language
At Lancaster University, United Kingdom
for the degree of
Doctor of Philosophy
November 2009

Abstract

This thesis conducted both quantitative and qualitative analyses of corpora of writing from L1 Chinese learners of L2 English and two native corpora, attempting to explore and identify the differences and similarities in the use of lexical bundles across learner proficiencies as well as between native and non-native writing.

The morphosyntactic features in second language writing have been extensively researched during the past decades. Few studies, however, have attempted to extend attention outward to the discourse aspect of learner writing by examining large quantities of empirical data. The present thesis hence addressed a textual perspective via a frequency-driven phraseological approach, i.e. to look into the discourse aspect of learner language development through *lexical bundles* (a.k.a. *recurrent word combinations*).

In Modular Study 1, learner essays written by L2 students were compared with two corpora of L1 written English: one referring to native expert writing and the other native peer writing. The native expert writing was extracted from the component of academic prose in the FLOB corpus (FLOB-J). The two groups of student writing, L2 writing of L1 Chinese students (BAWE-CH) and L1 peer writing of British students (BAWE-EN), both come from the BAWE corpus, which compiled proficient assessed student writing from British universities.

In Modular Study 2, argumentative and expository essays chosen from the Longman Learner Corpus were rated by at least two experienced raters. Adopting a rigorous rating procedure (including benchmarking, rater training, and statistic analyses) as generally used in high-stakes language tests, proficiency was determined with the Common European Framework of Reference (CEFR). Two sizeable subcorpora representing two CEFR levels, B2 and C1, were selected for investigation.

Through various ways of comparison, i.e. structural and functional categorisation as well as the keyness analysis, a few developmental patterns in the use of lexical bundles have been identified. The results show that at the lower proficiency levels, learner language tends to be more simplistic, colloquial, clichéd, verbose, categorical, and overstating. In comparison, the more proficient writing demonstrates an opposite pattern, thereby being more native-like in this regard. The interpretations of results and the implications for L2 writing pedagogy, language testing, and psycholinguistics will be discussed. A few methodological issues, such as the use of chi-square tests and determination of a frequency and dispersion threshold in bundle studies, will be addressed too.

Declaration

Many people have given me assistance in the course of this study, yet this thesis has been written by myself, and the work reported herein is entirely my own, which has not been previously submitted for a degree in this or any other form.

Yu-Hua Chen, November 2009

Abbreviation/Acronym List

BAWE: the British Academic Written English corpus

BNC: British National Corpus

CAE: Certificate in Advanced English, one of Cambridge ESOL general English exams

Cambridge ESOL Examinations: Cambridge Examinations of English for Speakers of Other Languages

CEFR: the Common European Framework of Reference

CPE: Certificate of Proficiency in English, one of Cambridge ESOL general English exams

CL: corpus linguistics

CLC: Cambridge Learner Corpus

EAP: English for Academic Purposes

ESP: English for Specific Purposes

KET: Key English Test, one of Cambridge ESOL general English exams

ELT: English Language Teaching

ESL: English as a second Language

EFL: English as a foreign language

FCE: First Certificate in English, one of Cambridge ESOL general English exams

FLOB: the Freiburg-LOB Corpus of British English corpus (also called the Freiburg-Lancaster-Oslo/Bergen corpus)

ICLE: the International Corpus of Learner English

IELTS: International English Language Testing System

L1: first language

L2: second language

LT: language testing

LLC: Longman Learners' Corpus

PET: Preliminary English Test, one of Cambridge ESOL general English exams

SLA: second language acquisition

TOEFL: Test of English as a Foreign Language

ABBREVIATION/ACRONYM LIST I
LIST OF TABLES V
LIST OF FIGURES IX
LIST OF EQUATIONS X
CHAPTER 1 INTRODUCTION 1
1.1 The Goal of this Thesis 1
1.2 Defining Key terms 1
1.3 Research Questions 5
1.4 Structure of Thesis 9
CHAPTER 2 RESEARCH BACKGROUND 12
2.1 Corpus Approaches and Language Learning 14
2.1.1 Corpus Linguistics 15
2.1.2 Learner Corpora 15
2.1.3 Corpus Research & Second/Foreign Language Learning 21
2.2 Developmental Studies and Proficiency Levels 27
2.3 Learner Proficiency in SLA, Learner Corpus Research, and Language Testing 30
2.4 Conclusion 33
CHAPTER 3 THEORETICAL AND OPERATIONAL FRAMEWORK 36
3.1 The Study of Phraseology: Word Co-occurrence 38
3.2 Lexical Bundles in L1 and L2 Studies 38
3.3 Determination of Lexical Bundles 42
3.4 Corpora Investigated 45
3.5 Automatic Retrieval 49
3.5.1 Determination of Frequency and Dispersion Thresholds 50
3.5.2 Results of Automated Retrieval 50
3.6 Manual Examination 56
3.6.1 Context Independence 57
3.6.2 Data Deflation 57
3.7 Conclusion 60
CHAPTER 4 ANALYTICAL FRAMEWORK 69
4.1 Ways of Comparing Corpora 70
4.1.1 Two Modular Studies 70
4.1.2 Types vs. Tokens 70
4.1.3 Keyness Analysis 73
4.1.4 Structural and Functional Distribution 74
4.2 Structural Classification 75
4.2.1 Background 76
4.2.2 Issues Concerning Structural Categorisation 76
4.2.3 A Modified System for Structural Classification 79
4.3 Functional Classification 87
4.3.1 Background 94
4.3.2 Issues Concerning Functional Categorisation 94
4.3.3 A Modified System for Functional Classification 97
4.4 Conclusion 99
105

CHAPTER 5 MODULAR STUDY 1: LEXICAL BUNDLES IN NATIVE WRITING AND LEARNER WRITING WITHIN THE ACADEMIC CONTEXT 106
5.1 Selection of Data 106
5.2 Linguistic Profile 106
5.2.1 Type/token Ratio 112
5.2.2 Words with Different Lengths 113
5.2.3 Topics & Titles of the Written Samples 114
5.3 Structural Analysis 115
5.3.1 Type Distribution 117
5.3.2 Token Distribution 118
5.3.3 The Chi-square Test and Standardised Residuals 119
5.3.4 Lexical Bundles in Structural Categories 121
5.4 Functional Analysis 126
5.4.1 Type Distribution 136
5.4.2 Token Distribution 136
5.4.3 The Chi-square Test & Standardised Residuals 138
5.4.4 Lexical Bundles in Functional Categories 140
5.5 Relationship between Structural and Functional Categorisation 143
5.6 Keyness Analysis 160
5.7 Summary of Findings 162
CHAPTER 6 MODULAR STUDY 2: LEXICAL BUNDLES ACROSS LEARNER PROFICIENCIES 165
6.1 Determination of Proficiency 168
6.1.1 Benchmarking 169
6.1.2 Major Rating 170
6.2 Selection of Data 185
6.3 Linguistic Profile 195
6.3.1 Learner Backgrounds 198
6.3.2 Type/token Ratio 198
6.3.3 Words with Different Lengths 199
6.3.4 Topics of the Written Samples 201
6.4 Structural Analysis 202
6.4.1 Type Distribution 203
6.4.2 Token Distribution 203
6.4.3 The Chi-square Test and Standardised Residuals 205
6.4.4 Use of Lexical Bundles in Structural Categories 206
6.5 Functional Analysis 209
6.5.1 Type Distribution 215
6.5.2 Token Distribution 215
6.5.3 The Chi-square Test and Standardised Residuals 217
6.5.4 Use of Lexical Bundles in Functional Categories 219
6.6 Relationship between Structural and Functional Categorisation 221
6.7 Keyness Analysis 234
6.8 Summary of Findings 236
..... 238

CHAPTER 7 OVERVIEW & DISCUSSION **241**
7.1 Comparison of Analyses **241**
7.1.1 Quantitative Analyses 242
7.1.2 Qualitative Examination of Discourse Functions 246
7.1.3 Good Indicators for Writing Competency 259
7.1.4 Keyness Analysis 263
7.1.5 Overall Development & Some Methodological Issues 266
7.2 Distinctive Features across Proficiencies **277**
7.2.1 Towards a More Complex Language 278
7.2.2 Cliché & Verbosity vs. Concise Language 280
7.2.3 Colloquialism vs. Formality 283
7.2.4 Overgeneralisation vs. Cautious Language 284
7.2.5 Shift of Authorial Voice 287
7.3 'Foreign-soundingness' in the Use of Lexical Bundles **288**
7.4 Summary **298**
CHAPTER 8 CONCLUSION **300**
8.1 Status of Frequency-Driven Formulaic Language in SLA **300**
8.2 Limitations **303**
8.3 Implications **306**
8.3.1 Methodological Issues 306
8.3.2 Psycholinguistics 310
8.3.3 Second Language Teaching & Learning 313
8.3.4 SLA & Language Testing 315
8.4 Future Research Directions **316**
8.5 Concluding Remarks **320**
APPENDIX 1 PECULIAR CONDITIONS OF OVERLAPPING BUNDLES **322**
APPENDIX 2 COMMON REFERENCE LEVELS: GLOBAL SCALE **326**
APPENDIX 3 CEFR WRITTEN ASSESSMENT CRITERIA GRID **327**
APPENDIX 4 LEXICAL BUNDLES IN FREQUENCY ORDER **328**
REFERENCES **332**

List of Tables

Table 2-1 Previous studies based on learner corpora of exam scripts from Cambridge ESOL (cited from Banerjee et al. (2007) with modification) 35
Table 3-1 Past studies on lexical bundles in written or written and spoken registers 47
Table 3-2 Past studies on lexical bundles in spoken registers only 48
Table 3-3 Threshold of raw frequency of four times occurring in at least three texts 52
Table 3-4 Threshold of normalised frequency of 20 times per million words occurring in at least three texts 52
Table 3-5 Raw and converted normalised frequency thresholds adopted 54
Table 3-6 Normalised and converted raw frequency thresholds for comparison 54
Table 3-7 Corpora of different sizes and the number of the retrieved lexical bundles 56
Table 3-8 Corpora used in two modular studies, the threshold, and the number of retrieved lexical bundles 57
Table 3-9 Number of lexical bundles before and after filtering out context-dependent bundles in Modular Study 1 (types) 59
Table 3-10 Number of lexical bundles before and after filtering out context-dependence in Modular Study 2 (types) 59
Table 3-11 Instances of 'Complete Overlaps' 62
Table 3-12 Instances of 'Complete Subsumption' 63
Table 3-13 Instances of 'Partial Subsumption' 65
Table 3-14 Number of lexical bundles before and after filtering and deflation in the corpora investigated (types) 68
Table 3-15 Number of bundles before and after the removal of context dependent bundles and overlaps 68
Table 4-1 Proportional distribution of four-word lexical bundles across the major structural patterns in LSWE, FLOB-J, BAWE-EN, and BAWE-CH (adapted from the Longman Spoken and Written English (Biber et al., 1999, p. 996, with the addition of the three academic corpora used in this thesis) 78
Table 4-2 Some examples of composite lexical bundles and their structural units 82
Table 4-3 Collocation and frequency in *at the end of* and *the end of the* 84
Table 4-4 Development of taxonomy for functional categorisation in studies conducted by Biber et al. 95
Table 4-5 Expressions incorporating *to a lack of* and their frequency in BAWE-EN 98
Table 5-1 Number of BAWE samples from L1 Chinese students who received different years of secondary education in the UK 108
Table 5-2 Number of samples written by students from different level of study in BAWE-CH and BAWE-EN 109
Table 5-3 Number of samples awarded the grade of merit (M) or distinction (D) in BAWE-CH and BAWE-EN 109

Table 5-4 Number of samples from different genres in BAWE-CH and BAWE-EN 110
Table 5-5 Number of samples from different disciplinary areas (Arts and Humanities (AH), Life Sciences (LS), Physical Sciences (PS), Social Sciences (SS)) 111
Table 5-6 Constituents of the three academic corpora 112
Table 5-7 STTR in Modular Study 1 113
Table 5-8 Numbers of words with different lengths in Modular Study 1 (occurrences) 115
Table 5-9 Some examples of essay/paper topics and book/chapter/paper titles in Modular Study 1 116
Table 5-10 Structural distribution in Modular Study 1 (types) 118
Table 5-11 Structural distribution in Modular Study 1 (tokens) 120
Table 5-12 Chi-square standardised residuals for structural distribution (types) in Modular Study 1 123
Table 5-13 Chi-square standardised residuals for structural distribution (tokens) in Modular Study 1 124
Table 5-14 NPf bundles shared by FLOB-J and BAWE-EN 128
Table 5-15 The frame '*the* + Noun + *of the/a*' used in Modular Study 1 128
Table 5-16 The frame '*in the* + Noun + *of* used in Modular Study 1 129
Table 5-17 Frequency of the orthographic forms *use/used* in Modular Study 1 131
Table 5-18 The subcategory of 'Pronoun/Noun + BE/Verb phrases' in Modular Study 1 132
Table 5-19 The subcategory of '*to*-clause fragments' in Modular Study 1 132
Table 5-20 Frequency of *in order to* in Modular Study 1 134
Table 5-21 The subcategory of 'Passive verb +prepositional phrases' in Modular Study 1 134
Table 5-22 Passive-verb bundles in the subcategory of 'VP + *to*-clause' in Modular Study 1 135
Table 5-23 Passive-verb bundles in the subcategory of '*it* pattern' in Modular Study 1 135
Table 5-24 Functional distribution in Modular Study 1 (types) 137
Table 5-25 Functional distribution in Modular Study 1 (tokens) 138
Table 5-26 Chi-square standardised residuals for functional distribution (types) in Modular Study 1 140
Table 5-27 Chi-square standardised residuals for functional distribution (tokens) in Modular Study 1 141
Table 5-28 Referential framing bundles in Modular Study 1 145
Table 5-29 Referential quantifying bundles in Modular Study 1 146
Table 5-30 Referential deictic bundles in Modular Study 1 147
Table 5-31 Stance epistemic bundles in Modular Study 1 152
Table 5-32 Stance attitudinal/modality bundles in Modular Study 1 154
Table 5-33 Topic elaboration/clarification bundles in Modular Study 1 156
Table 5-34 Frequency of *be seen as/be regarded as* in Modular Study 1 159
Table 5-35 Topic introduction bundles in Modular Study 1 159
Table 5-36 Identification/focus bundles in Modular Study 1 159
Table 5-37 Inferential bundles in Modular Study 1 160
Table 5-38 Key lexical bundles in BAWE-EN and BAWE-CH with FLOB-J as the reference corpus. 163
Table 5-39 Key lexical bundles in BAWE-CH with BAWE-EN as the reference corpus 165
Table 6-1 Transcription of CEFR levels to numerals 173

Table 6-2 Descriptive statistics for benchmarking 173
Table 6-3 Correlation coefficients for benchmarking 174
Table 6-4 All Facet vertical 'rulers' for benchmarking 178
Table 6-5 Candidate measurement report for benchmarking 180
Table 6-6 Instance of misfitting #12 in benchmarking 181
Table 6-7 Rater measurement report for benchmarking 182
Table 6-8 Unexpected responses generated by FACETS 183
Table 6-9 Unexpected marking: #20 183
Table 6-10 Descriptive statistics for major rating 188
Table 6-11 Inter-rater reliability for major rating 189
Table 6-12 All Facet vertical 'rulers' for major rating 190
Table 6-13 Rater Measurement Report 191
Table 6-14 Candidate Measurement Report 191
Table 6-15 Corpus size and number of texts built on the basis of Rasch-calibrated ratings 194
Table 6-16 Corpus size and number of texts of the original rated data and the selected rated data . 197
Table 6-17 STTR in Modular Study 2 200
Table 6-18 Reproduction of Table 5-7 (STTR in Modular Study 1) 200
Table 6-19 Numbers of words with different lengths in Modular Study 2 (occurrences) 201
Table 6-20 Some examples of essay topics in Modular Study 2 202
Table 6-21 Structural distribution in Modular Study 2 (types) 203
Table 6-22 Structural distribution in Modular Study 2 (tokens) 205
Table 6-23 Chi-square standardised residuals for structural distribution (types) in Modular Study 2 208
Table 6-24 Chi-square standardised residuals for structural distribution (tokens) in Modular Study 2 208
Table 6-25 The frame '*the* + Noun + *of the/a*' used in Modular Study 2 211
Table 6-26 Lexical bundles and the frequency in the subcategory of PPf bundles in Modular Studies 1 & 2 212
Table 6-27 Numbers of bundles in VP-based subcategories 213
Table 6-28 Bundles with passive verbs in VP-based bundles in Modular Study 2 215
Table 6-29 Functional distribution in Modular Study 2 (types) 216
Table 6-30 Functional distribution in Modular Study 2 (tokens) 217
Table 6-31 Chi-square standardised residuals for functional distribution (types) in Modular Study 2 219
Table 6-32 Chi-square standardised residuals for functional distribution (tokens) in Modular Study 2 220
Table 6-33 Referential quantifying bundles in Modular Study 2 223
Table 6-34 Referential framing bundles in Modular Study 2 225
Table 6-35 Referential deictic bundles in Modular Study 2 225
Table 6-36 Stance epistemic bundles in Modular Study 2 226
Table 6-37 Stance attitudinal/modality bundles in Modular Study 2 229
Table 6-38 Topic introduction bundles in Modular Study 2 230

Table 6-39 Collocation after the topic-introduction bundle in CEFR-B2 and CEFR-C1 writing 231
Table 6-40 Topic elaboration/clarification bundles in Modular Study 2 233
Table 6-41 Inferential bundles in Modular Study 2 233
Table 6-42 Identification/focus bundles in Modular Study 2 234
Table 6-43 Key lexical bundles in CEFR-B2 with CEFR-C1 as the reference corpus 236
Table 7-1 Referential quantifying bundles in CEFR-B2 248
Table 7-2 Epistemic bundles with the gradient of certainty 252
Table 7-3 Use of first person pronouns in lexical bundles 253
Table 7-4 Inferential bundles in Modular Studies 1 and 2 258
Table 7-5 Overall chi-square standardised residuals in structural distribution (types) 259
Table 7-6 Overall chi-square standardised residuals in structural distribution (tokens) 260
Table 7-7 Overall chi-square standardised residuals in functional distribution (types) 261
Table 7-8 Overall chi-square standardised residuals in structural distribution (tokens) 261
Table 7-9 Key bundles shared in student and learner corpora with FLOB-J as the reference corpus 264
Table 7-10 Constituents of corpora and cut-off frequency for determining lexical bundles 271
Table 7-11 Bundle types and tokens before and after normalisation (per 100,000 words) 272
Table 7-12 Spread of lexical bundles (tokens) 273
Table 7-13 Proportion of formulaicity reported in the literature 275
Table 7-14 Clichéd bundles 281
Table 7-15 A selection of overgeneralising bundles 286
Table 8-1 Lexical bundles in CEFR-B1 writing (bundle criteria altered to 3 times in 2 texts) 318
Table 0-1 Peculiar overlapping bundles in Modular Study 1 324
Table 0-2 Peculiar overlapping bundles in Modular Study 2 325

List of Figures

Figure 2-1 A wordlist with four-word clusters ordered by frequency 16
Figure 2-2 KWIC with the example of *on the other hand* 18
Figure 2-3 Collocation with the search word *way* 19
Figure 4-1 Shifting categorisation of certain types of bundles 96
Figure 5-1 Structural distribution of lexical bundles in Modular Study 1 (types) 119
Figure 5-2 Structural distribution in Modular Study 1 (tokens) 120
Figure 5-3 Functional distribution in Modular Study 1 (types) 137
Figure 5-4 Functional distribution in Modular Study 1 (tokens) 139
Figure 5-5 Breakdown of referential expression in Modular Study 1 (types) 144
Figure 5-6 Breakdown of stance bundles in Modular Study 1 (types) 150
Figure 5-7 Proportional use of breakdown of stances bundles in Modular Study 1 (types) 151
Figure 5-8 Breakdown of discourse organisers in Modular Study 1 (types) 155
Figure 5-9 Interaction between structural and functional categorisation in Modular Study 1 161
Figure 6-1 Visual representation of standardisation of judgments (extracted and modified from Figure 1.1 Section A in *Reference supplement to the preliminary pilot version of the Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2004)) 170
Figure 6-2 Text types in the finalised CEFR-B2 corpus and CEFR-C1 corpus 197
Figure 6-3 Learner backgrounds in the CEFR-B2 corpus and CEFR-C1 corpus 199
Figure 6-4 Structural distribution of lexical bundles in Modular Study 2 (types) 204
Figure 6-5 Structural distribution in Modular Study 2 (tokens) 206
Figure 6-6 Functional distribution in Modular Study 2 (types) 216
Figure 6-7 Functional distribution in Modular Study 2 (tokens) 218
Figure 6-8 Breakdown of referential expressions in Modular Study 2 (types) 222
Figure 6-9 Breakdown of stance bundles in Modular Study 2 (types) 226
Figure 6-10 Breakdown of discourse organisers in Modular Study 2 (types) 230
Figure 6-11 Interaction between structural and functional categorisation in Modular Study 2 235
Figure 7-1 Overall structural distribution (types-percentage) 243
Figure 7-2 Overall structural distribution (tokens-percentage) 243
Figure 7-3 Overall functional distribution (types-percentage) 245
Figure 7-4 Overall functional distribution (tokens-percentage) 245
Figure 7-5 Distribution of referential subcategories (types-percentage) 247
Figure 7-6 Distribution of referential subcategories (tokens-percentage) 247
Figure 7-7 Distribution of stance subcategories (types-percentage) 250
Figure 7-8 Distribution of stance subcategories (tokens-percentage) 250
Figure 7-9 Distribution of discourse organising subcategories (types-percentage) 255
Figure 7-10 Distribution of discourse organising subcategories (tokens-percentage) 256
Figure 7-11 Overall number of bundle types and tokens per 100,000 words 267
Figure 7-12 Distribution of structural subcategories (types) 279
Figure 7-13 Relative frequency of *on the other hand* and *at the same time* 290
Figure 7-14 Relative frequency of *in terms of the* and *in the case of* 291

List of Equations

Equation 5-1 Formula of chi-square statistic χ^2 121
Equation 5-2 Formula of a standardised residual 122

Chapter 1 Introduction

This chapter summarises the core notions in this thesis. The first section starts with the goals of this research, and the next section moves on to the introduction of key terms used throughout this thesis. Then the research questions, grouped according to the subject areas involved, will be addressed. This chapter will end with an overview outlining the structure of this thesis.

1.1 The Goals of this Thesis

This thesis aims to conduct both quantitative and qualitative analyses of learner corpora from L1 Chinese learners of L2 English and two L1 English corpora, with the aim to explore and identify the similarities and differences in the use of recurrent word combinations between L1 and L2 writing as well as across L2 proficiencies. Learners' language development is generally described and analysed in terms of fluency, accuracy and complexity. Few studies, however, have attempted to extend attention outward to the discourse aspect of second language development by examining large quantities of empirical data. This study hence intends to take a textual perspective via corpus approaches, i.e. to look into learner language through recurrent word combinations (a.k.a. *lexical bundles*). Recurrent word combinations are computer-derived phraseological units, which are defined with a specified frequency and distribution threshold and have been found to often function as the 'building blocks' of discourse. This is therefore a frequency-driven approach which works on the discourse aspect of learner language from a phraseological perspective, instead of extensively researched morphosyntactic structures in second language research.

In the past few decades, researchers have become increasingly interested in how

words co-occur in discourse to form formulaic units (e.g. *things like that*, *pay attention*, *in the context of*, *as well as*). With the development of corpus linguistics, i.e. the study of language patterns through collections of machine-readable texts, some recent studies have added more weight to the significance of multi-word expressions in language acquisition on the basis of empirical evidence established upon corpus data. The investigation of learner writing and native writing in this thesis is thus a study which focuses on recurrent strings of continuous word co-occurrence, using both 'corpus-driven' and 'corpus-based' methods (for a detailed comparison of the corpus-driven and the corpus-based approaches, please see Tognini-Bonelli, 2001). Without any preconceptions about linguistic forms or functions, a list of uninterrupted word sequences, along with their frequencies, are retrieved from the corpora—which is a bottom-up and 'corpus-driven' approach. Then a set of structural and functional taxonomies developed by Biber and his colleagues (e.g. 1999, 2003, 2004) are adopted so as to classify those computer-derived word sequences—which is a top-down and 'corpus-based' approach. For the sake of comparability with the literature, only 4-word combinations are investigated in this thesis as they have been the most researched length of recurrent word combinations.

In order to allow for comparisons to be made between native writing and learner writing as well as across learners' proficiency levels, two modular studies are designed with the use of different corpora dealing with different genres of writing. The first modular study (described in Chapter 5) compares L2 English writing from L1 Chinese students in British higher education with L1 English peer student writing and L1 English published academic prose. The second modular study (described in Chapter 6) compares L1 Chinese learner of L2 English writing for academic purposes, either argumentative or expository essays, between two proficiency levels defined with the Common European Framework of Reference (CEFR), CEFR-B2 and CEFR-C1. From the first comparative study, we can know to what extent the L2 learners in the British higher education have approximated native standards. Can learner

performance go beyond the performance of the native peers (the British students) and approach the norm of native professional writers? Or if combined with the results of the second comparative study, do the learners across proficiency levels share some common textual features and as a whole appear significantly distinct from native writers with respect to their use of different types of lexical bundles? The overall developmental patterns and the possible explanations for the results will be given in Chapter 7. Through comparisons across various learner groups and native groups as well as consulting other sources of information such as coursebooks on EAP (English for Academic Purposes), it is hoped that this thesis can throw light on a better understanding of learner language.

This thesis also deals with a couple of methodological issues. In terms of defining lexical bundles, the interaction between corpus size and cut-off frequency and distribution is found to be more complex than expected, particularly as the corpora used in this thesis are not of equal size and are rather small in comparison with most corpus studies of native-speaker English. In addition, the recurrent word sequences extracted from the automated procedure are not suitable for analysis until undesired 'noise' such as overlapping or context-dependent word sequences are manually filtered out, a finding which surprisingly has rarely been reported in the literature. With regard to the categorisation of lexical bundles, the greatest challenge lies in the fact that the assignment of a corresponding category can be ambiguous and controversial, suggesting the lack of clear-cut demarcation in such categorisation. Referring to the literature, it is found that sometimes certain bundles are categorised as having one function and sometimes as having another. Even the categorisation framework seems to shift in various studies. These challenges and their suggested solutions will be addressed in this thesis. The final methodological issue involves the construction of learner corpora and relates to the determination of proficiency. In past studies of second language acquisition, the definition of proficiency levels is usually vague, subjective and

sometimes not sufficiently fine-grained. Researchers often resort to 'extra-linguistic' judgments, e.g. years of learning English, instead of the written performance's linguistic features (Atkins & Clear, 1992, p. 5; Granger, 1998a, p. 9). The current project argues that a well-designed rating procedure, as generally followed in large-scale language tests, should be adopted to decide each script's proficiency level before researchers can accordingly compare learners' performance at different levels. The execution of such a procedure in the second modular study also proves to be an effective measure of determining learner proficiency in second language research.

There are two reasons why the study of lexical bundles can contribute to the area of English Language Teaching (ELT). On the one hand, words are traditionally the basic units in the vocabulary list and sentence constructions are the elementary grammatical patterns that learners are expected to acquire in order to master a foreign language. Research on lexical bundles, however, has suggested that these highly frequent multi-word expressions, many of which are lexico-syntactic units (e.g. *at the end of*, *the way in which*, *it is possible that*), have blurred the boundary between lexis and syntax while they serve as the building blocks of discourse (Biber & Conrad, 1999; Biber et al. 2003). On the other hand, although the importance of frequency has been recognised in vocabulary learning nowadays, quantitative phraseological data has shown that actually many words are highly frequent because they form components in many frequent fixed expressions (Stubbs, 2007a). Yet the above facts revealed by phraseological corpus studies during the past few decades do not appear to have inspired ELT publishers or practitioners to place more emphasis on (computer-retrieved) formulaic language in their curricula or materials. Through an examination of the literature and the researcher's investigation of both native and non-native corpora, coupled with the review of a number of ELT materials, this thesis argues that phraseology plays an important role as well as vocabulary in acquiring a language and that the definition of 'vocabulary list'

should be reinterpreted. It is suggested that those frequency-driven phraseological items, after proper editing and selection, should be incorporated into essential vocabulary lists in the future. It is also hoped that such a corpus-oriented investigation of recurrent word combinations in learner and native writing will not only provide insights for second language pedagogy but also shed light on learners' language development. Meanwhile, the analysis results may facilitate the advancement of language testing research, e.g. developing a computer automated marking system or a common band scale of writing assessment in the future.

For the remainder of this introductory chapter, I provide definitions of some of the most frequently used terms in this thesis, describe the research questions that I aim to address, and also outline the overall structure of the thesis.

1.2 Defining Key terms

For the purpose of clarity, a number of key terms will be addressed and explained here in alphabetical order, although some of them will be further discussed later in the thesis.

- **CEFR**: CEFR stands for the Common European Framework of Reference, which is a reference framework describing six proficiency levels (from the breakthrough level A1, then A2, B1, B2, C1, to the most proficient level C2) in second language learning. Developed by a group of experts under the administration of the Language Policy Division of the Council of Europe, the CEFR had gone through an extensive range of research and consultation. This framework 'provides a basis for the mutual recognition of language qualifications' between the citizens of EU member states (see the official Council of Europe website: http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp) and has been widely used in various language related areas such as language education or language testing (Council of Europe, 2003).
- **ESL vs. EFL**: some researchers distinguish ESL (English as a Second Language) and

EFL (English as a Foreign Language) while others do not. The former refers to learning of English as the non-mother-tongue language in an English-speaking country and the latter in a non-English-speaking country. This distinction is not emphasised here although most of the data used in this study could be considered under the EFL remit as most L1 Chinese-speaking students are known to acquire their L2 English in their home country. Generally speaking, this thesis discusses the general conditions in which English is taught and learnt as the target language other than the learners' mother tongue Chinese, regardless of the geographical area where teaching and learning takes place. In addition, the language produced by second/foreign language learners in the process of acquiring the target language can be termed as interlanguage (Selinker, 1972).

- **Idiomaticity vs. compositionality**: idiomatic expressions are the phraseological units which are often treated as holistic items rather than compositional ones. The determination of idiomaticity generally involves the notion of compositionality, i.e. whether the meaning consists of the sum of constituent elements in the word combinations (cf. Biber et al.1999, p. 1024; Moon, 1998a; Stubbs, 2002, p. 221). The meaning of the idiom *kick the bucket* ('to die'), for example, can not be derived from the literal meaning of the three constituent words (*kick* + *the* + *bucket*). In such a case, it is considered to be an idiomatic expression (other examples include *fall in love*, *beat around the bush*, *on the other hand*, *a piece of cake*). It should be noted that the notion of idiomaticity can involve different degrees of compositionality (Svensson, 2008). Figurative idioms such as *do a U-turn* still preserve their literal meaning despite their figurative interpretation while pure idioms such as *spill the beans* are by no means compositional when they are not treated as idioms (Cowie, 1981, 1998a; Granger & Paquot, 2008).

- **Keyness**: keyword analysis is a function provided by *WordSmith* (Scott, 2007), which identifies the words which are significantly more or less frequent in a target corpus when compared with a reference corpus. *WordSmith* performs statistical tests (the chi-square test or the log-likelihood test) which compare the frequencies of a word in both corpora, taking into account the overall size of each. This programme can also be performed on lexical bundles (called *clusters* in *WordSmith*) or on word class tags (e.g. the frequency of nouns in a corpus). Since the current study only deals with lexical bundles rather than individual words, the term 'keyness analysis' is hence coined to refer to this approach so as to avoid confusion with the commonly known 'keyword analysis'.
- **L1 vs. L2**: L1 is the first language, i.e. a person's mother tongue while L2 is the second language, which generally refers to any language(s) learned or acquired after L1. The L2 learners who contributed to the corpus data investigated in this thesis are L1 Chinese learners of L2 English regardless of their nationality or geographical origin (China, Taiwan or Hong Kong). Occasionally the terms 'native' and 'non-native' are also used interchangeably for L1 and L2.
- **Lexical bundles vs. recurrent word combinations**: lexical bundles and recurrent word combinations are used interchangeably in this thesis. The term 'lexical bundles' was first proposed by Biber and his colleagues (Biber & Conrad, 1999; Biber, et al., 1999), which refers to continuous word sequences occurring over a specified frequency and distribution threshold. The term 'recurrent word combinations' (Altenberg, 1998; De Cock, 1998), however, does not imply such a strict frequency and dispersion requirement. Instead, from its literal sense, it simply refers to the word combinations that occur more than once in the text in question and thus appears to be a more loosely defined term. The same frequency-driven approach has also been termed

differently in the literature, such as phrasicon (De Cock et al.1998), statistical phrases (Strzalkowski, 1998), chains (Stubbs, 2002), clusters (Scott, 2007), or n-grams (Fletcher, 2008; Stubbs, 2007a, 2007b). By and large, the word combinations retrieved with this approach can be regarded as one kind of phraseological units (see below for the definition of phraseology). A set of detailed definitions of 'lexical bundles' will be further discussed in Chapter 3.

- **Phraseology vs. formulaic sequences/language**: 'phraseology' (Cowie, 1998b; Granger & Meunier, 2008; Meunier & Granger, 2007) and 'formulaic sequences/language' (Schmitt, 2004; Wray, 2002, 2008) are two umbrella terms often used for word co-occurrence in fixed orders, both of which will be used interchangeably in this thesis. As Wray and Perkins have pointed out (2000, p. 3), over 40 terms have been used to describe the phenomenon of word combinations. On the one hand, various terms are used to refer to similar or even the same notion of word co-occurrence. On the other hand, the same term might be used in different ways by different scholars. In the broadest sense, phraseology or formulaic language can entail the most structurally fixed and semantically opaque form of word co-occurrence such as pure/frozen idioms (e.g. *kick the bucket*) to the most variable and transparent form such as free combination (e.g. *kick the football*) (for the phraseological continuum, see Cowie, 1981, 1998a). Phraseology and formulaic language both appear to have a basic assumption that these patterned linguistic units are stored and processed holistically as opposed to analytically in our mental lexicon; however, thus far this claim has still remained inconclusive in the area of psycholinguistics. This issue will be discussed in Chapter 3 and Chapter 7.

- **Types vs. tokens**: Types and tokens are commonly distinguished in linguistics. Types are the different word forms in a text while tokens simply refer to all the words in a text. For example, a text might have 100 words ('tokens'). Some of the 100 words, however, have identical word forms, i.e. repeating themselves (for example, the word 'elephant' counts as one type of word but may occur 10 times altogether), while others do not. All the words with different forms in this text are hence called 'types'. In this thesis, types and tokens are used to refer to recurrent word combinations rather than single words, unless specified otherwise.

In the following section, I will move on to address the research questions which provide the analytical focus for this thesis.

1.3 Research Questions

In a series of studies on lexical bundles conducted by Biber and his colleagues (e.g. Biber et al. 1999, 2003, 2004), it has been found that conversation and academic prose present distinctive types of distribution of lexical bundles. With regard to structural analysis, most bundles in conversation are clausal whereas most bundles in academic prose are phrasal.[1] When it comes to functional classification, the bundles retrieved from speech are primarily used as stance expressions (to express modality or attitude such as *I don't think so*) or interactional markers (which orient to the listener such as *you know what*). In writing, there are more referential bundles (which reference specific attributes such as *in the context of*) and

[1] It appears that 'phrasal bundles' refer to noun and prepositional word combinations while 'clausal bundles' are those with a verb component in Biber et al. (2004). There are, however, verb phrases as well as noun and prepositional phrases. The terms used to distinguish between phrasal and clausal bundles by Biber et al., therefore, can be ambiguous. This controversial issue will be further discussed in Section 4.2 Structural Classification.

discourse organisers (which introduce or clarify topics such as *as a result of*). Drawing on the previous framework, the current study intends to investigate whether there is a near-linear pattern in the distribution of bundle structures and functions across proficiency levels. That is to say, the general assumption being attested throughout the study is that the more proficient learners are, the closer their performance will approach native professional writing, i.e. being more register-sensitive and thus containing more noun-based and preposition-based bundles in terms of structures and more referential bundles and discourse organisers in terms of functions.

As can be seen, the current study is primarily purported to disclose the developmental patterns of lexical bundles across learners' writing proficiencies and the difference between learner writing and native writing. During the process of determining lexical bundles and categorising them, however, some methodological issues concerning the nature of those recurrent word combinations have arisen, and the scope of research hence extends further to this procedural aspect. Taking into account of the applications of analysis results involved in second language testing and instruction, the research questions are accordingly grouped into three dimensions: methodological questions, analytical questions and explanatory questions as below, and the abbreviations in the brackets indicate the related area (CL for Corpus Linguistics, SLA for Second Language Acquisition, and LT for Language Testing).

Methodological/Procedural questions

1. What are the optimum thresholds and procedures in terms of a) corpus size and b) frequency of bundles when investigating lexical bundle usage? (CL)
2. What problems are there with Biber et al's (1991, 2003, 2004, 2007) taxonomy for classifying bundle structures and functions, and how can the taxonomy be improved in order to create a consistent and robust categorisation scheme? (CL)

3. What is the most effective way of differentiating between learner proficiency levels? (SLA/LT)

Analytical questions

1. How do lexical bundles in learner performance differ from native language in terms of structures and functions? (SLA)
2. What is the developmental pattern in the written performance of L1 Chinese learners of L2 English in terms of their use of lexical bundles? (SLA)
3. What does a keyness analysis reveal about development of learner writing and learner writing versus native writing? (SLA/CL)

Explanatory questions

1. What are the possible reasons that result in the differences of use of lexical bundles in various target writers? (SLA)
2. What possible impacts do the results have with respect to improving pedagogy for second language writing? (SLA)
3. What possible impacts do the results have upon language testing, particularly upon the empirical underpinning of rating scales? (SLA/LT)

The present study is a collaboration between corpus linguistics, language testing, and second language research. If the above research questions are regrouped according to the three areas, it can be found that corpus linguistics and language testing basically bear on the methodological questions whilst the ultimate purpose of the research is chiefly concerned with second language acquisition. In other words, without solving the procedural questions relating to corpus linguistics and language testing, it is not possible to answer the analytical and explanatory questions with respect to second language acquisition. It is hoped that such an interdisciplinary collaboration can provide SLA research with more empirical descriptions as to learners' language development than have not been afforded in the past.

1.4 Structure of Thesis

This thesis has three major parts. The first part pinpoints the background of the thesis. This introductory chapter has addressed the aims of this thesis, definitions of a few key terms, and research questions. Chapter 2 is an overview of the rationale with regard to the main issues involved in this thesis, including corpus approaches and learner corpus studies, second language development, and second language learning. Definitions of learner proficiency from SLA perspectives and learner corpus research will also be discussed in this review chapter. The following framework section is divided into two chapters, Chapters 3 and 4. Chapter 3 focuses on lexical bundles as the theoretical and operational framework with which this thesis is built upon. This chapter first gives a critical review of the literature on lexical bundles and then introduces the corpora investigated in this thesis. What follows is a description of the methodological procedure of defining lexical bundles, which starts with the automatic retrieval of clusters and then moves on to manual examination to filter out overlaps and context-dependent bundles. After the bundles for investigation are finalised, Chapter 4 touches upon the analytical framework of this thesis. The ways of conducting and presenting the analyses are outlined in this chapter; meanwhile, how the current project has applied and modified Biber et al's taxonomy of structural and functional categorisation to the bundle data are also described, along with justifying the modifications made.

The second major part of this thesis reports the analysis results in the two modular studies of Chapters 5 and 6. Chapter 5 compares L2 learner writing with native expert writing and native peer student writing within the academic context. Chapter 6 compares L2 argumentative and expository essays at two CEFR levels, C1 and B2. At the beginning of Chapters 5 and 6, details are given with regard to the selection of corpus data along with its ethnographical and linguistic information. Chapter 6 particularly describes how the rating of proficiency levels was carried out, detailing undergoing various stages such as benchmarking,

rater training, and statistical analysis of the ratings.

This thesis ends with the third part, which consists of discussion and conclusion. Chapter 7 summarises the overall patterns found in the two modular studies by comparing the analysis results from the previous two chapters and also reports the possible interpretation of results. This chapter also discusses a number of discourse features which are found to be distinctive across writing proficiency levels. Chapter 8 addresses the status of the frequency-driven phraseology in SLA and also critically reviews the strengths and weaknesses of the approaches adopted in this thesis. In addition, implications to the methodological issues, language processing, second language pedagogy and language testing will be discussed. The chapter concludes by proposing directions for future research along with some concluding remarks.

Chapter 2 Research Background

As discussed in Section 1.3 Research Questions, the scope of this study is interdisciplinary in terms of the areas it is associated with. In relation to its primary purpose, this is a piece of SLA research which aims to explore learners' use of phraseological units across proficiency levels and also compare them against native-speaker 'norms'. As far as the methodology is concerned, the phraseological units that this study investigates are elicited from the data via corpus techniques, and the proficiency levels in learner data are determined with a standard rating procedure generally adopted in the field of language testing. There has already been some work in the areas of lexical bundle research and learner corpus studies. Additionally, a number of studies have started to combine the methods in SLA and language testing such as those using test candidates' performance data from high-stakes exams[2] in order to identify distinguishing features across proficiency levels. With the goal of exploring the use of lexical bundles in learner writing and native writing, this thesis is an attempt to further the close integration of approaches from language testing and corpus linguistics.

In view of its complexity and substantial length, the theoretical, operational, and analytical frameworks of lexical bundles will be dealt with in Chapters 3 and 4. The current chapter will provide a background with regard to the integration of corpus approaches, second language acquisition (SLA) studies, and language testing. The development of research in these core areas will be briefly reviewed, and the key concept of proficiency determination in SLA will also be examined.

---

[2] High-stakes exams are those which usually have a great impact on test takers' futures, e.g. university entrance exams. On the contrary, low-stakes exams such as achievement tests or placement tests generally do not affect test takers' life decisions (Davies, et al., 1999, p. 185).

2.1 Corpus Approaches and Language Learning

2.1.1 Corpus Linguistics

In Latin, *corpus* refers to 'body'. In the context of linguistics, a corpus is used to refer to any body of text in its broadest sense. Yet nowadays the term *corpus* is generally reserved for a large set of selected text stored and processed in a computer. According to McEnery & Wilson (1996, pp. 29-32), a modern corpus has the four major characteristics as below:

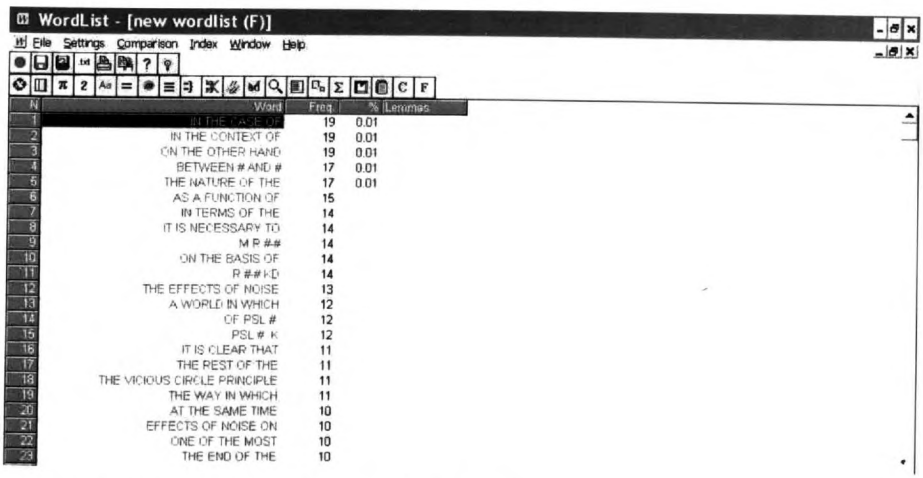
- sampling and representativeness;
- finite size;
- machine-readable form; and
- a standard reference.

Corpus linguistics, as some corpus linguists have pointed out (e.g. McEnery & Wilson, 1996, p. 2; Teubert & Krishnamurthy, 2007, p. 1), is a practice or a set of methodologies for the study of real life language rather than a branch of linguistics. In recent decades, corpus approaches have been applied to various areas in linguistics. The impacts these new approaches have made are revolutionary in the sense that corpus linguistics provides some empirical underpinnings incorporated with statistical measures which can deal with large amounts of linguistic data, which traditional descriptive linguistics could not afford.

A corpus, however, is simply an archive of collected texts if not processed by text retrieval software so that 'observations of various kinds can be made' (Hunston, 2002, p. 3). Most of the corpus tools currently available can easily allow researchers to sort the linguistic data into a specified type of order such as by frequency, by query, or by degree of significance when making comparisons. Take *WordSmith* (Scott, 2007), one of the most widely used corpus tools, for example. *WordSmith* allows the user to carry out the following analytical processes and information:

- wordlists and corresponding frequencies;
- lengths of whole scripts, paragraphs, sentences and words;
- lexical variation (type/token ratios);
- words in concordance and collocations; and
- keyword analysis.

The above features will be described here briefly as this thesis relies heavily on some of the approaches to analyse the data. First, the most fundamental information provided by a corpus software is the descriptive details regarding the constituents in a corpus: wordlists, corresponding frequency of each word, types (distinct words in text), tokens (running words in texts), and lengths of various units such as individual texts, paragraphs, sentences, and words. In *WordSmith*, the WordList function can also process continuous word strings with a specified length (called *clusters* in *WordSmith*) instead of single words, and this is the function being deployed by the current thesis (see Figure 2.1 for an example).



N	Word	Freq	% Lemmas
1	IN THE CASE OF	19	0.01
2	IN THE CONTEXT OF	19	0.01
3	ON THE OTHER HAND	19	0.01
4	BETWEEN # AND #	17	0.01
5	THE NATURE OF THE	17	0.01
6	AS A FUNCTION OF	15	
7	IN TERMS OF THE	14	
8	IT IS NECESSARY TO	14	
9	M R ##	14	
10	ON THE BASIS OF	14	
11	R ## K D	14	
12	THE EFFECTS OF NOISE	13	
13	A WORLD IN WHICH	12	
14	OF PSL #	12	
15	PSL # K	12	
16	IT IS CLEAR THAT	11	
17	THE REST OF THE	11	
18	THE VICIOUS CIRCLE PRINCIPLE	11	
19	THE WAY IN WHICH	11	
20	AT THE SAME TIME	10	
21	EFFECTS OF NOISE ON	10	
22	ONE OF THE MOST	10	
23	THE END OF THE	10	

Figure 2-1 A wordlist with four-word clusters ordered by frequency

In the statistic report of a wordlist computed by *WordSmith*, the type/token ratios for the overall corpus and individual texts are also provided. The type/token ratio is commonly used for comparing vocabulary complexity. It is believed that the higher the ratio, the more complex the language being investigated because a higher ratio means a wider range of vocabulary is included in the text. Yet type/token ratios are very sensitive to text length. If the corpus size increases, then the type/token ratio score would decrease as more words would reoccur (Wolfe-Quintero et al., 1998, p. 102). To remedy this problem, *WordSmith* computes the type/token ratios for every 1,000 tokens and then reports the average score.[3] The issue of type/token ratios will be touched upon later in Chapter 5 to illustrate the extent of lexical complexity in different groups of writing.

The next primary function corpus linguists often take advantage of is concordancing. A concordance shows all the occurrences of a search item, i.e. a word or a cluster. KWIC (key word in context)[4] is one of the most common formats for presenting the concordance listings (see Figure 2.2), which displays the search item in the middle and the context (which can be further expanded if needed) on both sides. This function facilitates a close examination of the environments where the items under investigation occur, thereby enabling the researchers to make generalisations from the patterns emerging from the environments.

---

[3] For learner writing analysed in this thesis, however, some of the texts are shorter than 1,000 tokens. In such cases, it is still unclear whether there is an impact on the way type/token ratios are computed in *WordSmith*.

[4] It has to be noted that the term *key word* used in concordance is completely different from the notion of *keyword* analysis. The former simply refers to the items under examination while the latter is reserved for the items highlighted through statistical measures.

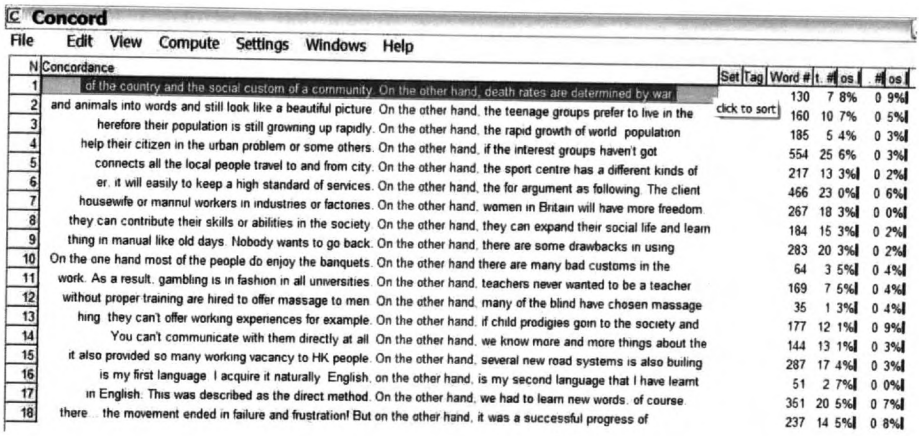


Figure 2-2 KWIC with the example of *on the other hand*

Similar to a concordance, collocation also pertains to the context of word occurrence. To be more specific, collocation refers to the relationship of certain words that co-occur close or next to each other. A corpus tool can easily compute the extent to which certain words co-occur and present the results in a systematic manner,[5] which allows researchers to better understand the meaning(s) and the usage of a word or a multi-word unit. Take Figure 2.3 for instance, with *way* as the search word (with 127 occurrences). We can see the words collocated with *way* within the range of five neighbouring words on both the right and the left sides, with the main collocates in the positions that occur most frequently highlighted in the colour red (L1, R1, for example). We thus know that *way* is most often collocated (in this particular corpus) with articles such as *the*, *this*, *one*, *some* preceding it and with the relative pronoun *which* following it. Collocation computed by a corpus tool is very useful when large amounts of data is processed with a great many concordance lines to be examined. For the current study, nevertheless, the highest number of occurrences for the four-word clusters

[5] Statistical measures such as Mutual Information (MI) score, t-score, z-score, or log-likelihood are commonly used to determine the significance of word co-occurrence.

investigated is only 36. It thus appears that concordancing alone suffices the purpose of checking the contexts for clusters and there seems no need to make use of collocation analysis for cluster research. On the other hand, lexical bundles (referred to also as *clusters* in *WordSmith*) could be thought of as a form of collocation, although while collocates generally can occur in different positions in relation to each other (e.g. *tell – story*), lexical bundles appear in fixed patterns (e.g. *the back of the*).

Concord - [collocates (total)]

C File View Settings Window Help

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	WAY	129	1	1	1	0	0	0	0	127	0	0	0	1	0
2	THE	104	60	44	8	7	3	19	23	0	9	11	6	8	10
3	IN	72	59	13	3	6	20	30	0	0	2	4	1	2	4
4	A	41	30	11	6	1	2	11	10	0	2	2	5	1	1
5	TO	37	19	18	7	5	3	4	0	0	5	4	3	2	4
6	OF	34	9	25	4	4	1	0	0	0	5	4	9	2	5
7	IS	28	12	16	0	5	7	0	0	0	2	3	4	4	3
8	THIS	25	20	5	2	3	1	3	11	0	0	2	0	1	2
9	WHICH	20	2	18	2	0	0	0	0	0	15	1	0	1	1
10	AND	17	9	8	3	2	4	0	0	0	0	2	3	1	2
11	ONE	13	7	6	0	1	0	0	6	0	1	1	2	0	2
12	BE	11	5	6	1	1	3	0	0	0	0	0	1	3	2
13	IT	11	4	7	0	2	0	2	0	0	2	1	2	2	0
14	NOT	10	4	6	1	2	1	0	0	0	0	0	2	3	1
15	SOME	10	8	2	0	0	1	1	6	0	0	0	0	2	0
16	THAT	9	3	6	0	0	1	0	2	0	2	1	2	1	0
17	AT	8	3	5	0	0	3	0	0	0	0	2	1	0	2
18	ANY	7	4	3	0	0	0	0	2	2	0	0	1	1	0	1
19	ARE	7	2	5	0	1	1	0	0	0	1	0	0	2	2
20	BY	7	4	3	1	0	0	1	2	0	1	0	1	1	0
21	ON	7	5	2	0	0	0	0	5	0	0	0	0	1	1
22	THEIR	7	4	3	1	1	1	0	1	0	0	2	0	1	0

Figure 2-3 Collocates of the search word *way*

Now turning to a keyword analysis, as introduced in Chapter 1, this procedure identifies the words which are significantly more or less frequent in a target corpus when compared against a reference corpus. As the current study examines multi-word units as opposed to words, the term 'keyness analysis' is hence coined to refer to this approach so as to avoid confusion with the commonly known 'keyword analysis'. More details about keyness analysis and its application in this thesis will be discussed in the following chapters.

The above are the common functions or analytical procedures that corpus software or an online interface to a corpus often provides. In order to facilitate more sophisticated

analyses, researchers can also annotate the raw corpus data either automatically or semi-automatically. Corpus annotation, or tagging, is 'the practice of adding interpretative (especially linguistic) information to an existing corpus by some kind of coding attached, or interpreted with, the electronic representation of the language material itself' (Leech, 1993, p. 275). There are several kinds of annotation: part-of-speech (POS) tagging, syntactic tagging (parsing), semantic tagging, discoursal tagging, and error tagging, and various software tools are available for different types of corpus annotation (see Granger, 2002 and Meunier, 1998 for a detailed summary of annotation tools). However, for learner corpora, a collection of texts produced by learners of a language, it can be expected that such annotation is particularly difficult in view of the idiosyncrasies and unpredictability in learner language in comparison with native language. Milton & Chowdhury's work (1994), to my knowledge, is one the earliest studies which raised the issue of learner errors which could greatly impact on corpus annotation. Error tagging, therefore, has to be carried out before making meaningful annotation of other forms in learner corpora. As error-tagging is extremely time and labour intensive, despite the existence of a few error editors, there have been few error-tagged learner corpora. The International Corpus of Learner English (ICLE) (Dagneaux et al. 1998) and the Cambridge Learner Corpus (Nicholls, 2003) are two of the pioneering error-tagged learner corpora. The systems of error tagging, usually a complicated multi-tiered one, have been reported in Granger (2003), Lüdeling et al. (2005), and Chuang & Nesi (2006). Yet the semi-automated frequency-driven approach adopted in this thesis, i.e. investigating the use of *lexical bundles*, will demonstrate that it is still possible to describe learner language without much annotation.

In the next section, we will see how second language research has benefited from corpus approaches and formed a newly emerging wave of learner corpus studies.

2.1.2 Learner Corpora

In comparison with prototypical native corpora, a learner corpus is a collection of spoken or written texts produced by second or foreign language learners. In the past, second language acquisition (SLA) research generally relied on a rather small quantity of empirical data, often on the basis of a number of subjects, 'which consequently raises questions about the generalisability of the results' (Granger, 2002, p. 6). In contrast with conventional SLA research, learner corpora tend to be much larger in terms of data size and the texts are contributed by a much wider range of learners. Just as with any practices of data collection in SLA research, however, compilation of a good learner corpus also involves careful consideration of variables including learner attributes and task settings. Learner attributes consist of information pertaining to learners' backgrounds such as mother tongue, proficiency, region, age, and so on, while task setting includes various settings relating to how the texts are generated such as whether they are naturalistic data or experimental data. Drawing on both corpus linguistics and SLA studies, learner corpus research takes advantage of the approaches of corpus linguistics for the purpose of a better understanding of second or foreign language learning.

One of the most well-known learner corpora is ICLE, the International Corpus of Learner English, which is composed of essays written by supposedly 'advanced' learners of English from various L1 backgrounds (Granger1993a; Granger1993b). At the time of writing, ICLE has accumulated data of over three million words from learners of 21 different L1 backgrounds, functioning as the source of numerous learner corpus studies (e.g. Cosme, 2006; Gilquin, 2002; Lorenz, 1998, to name just a few). Another large learner corpus, the Longman Learners' Corpus, is a ten-million-word computerised collection of learners' written English from various mother tongues and proficiency backgrounds, and this thesis used part of the Longman corpus data. One problem in the construction of these learner corpora that may

undermine the effects of analysis is the determination of learner proficiency. More often than not, the researchers resort to 'external criteria' rather than 'internal criteria' in this regard (Atkins & Clear, 1992, p. 5). 'Whilst proficiency level is of primary importance,' as Granger (1998a, p. 8) acknowledges, it is meanwhile a 'subjective notion' for most learner corpus compilations. This issue of applying external criteria such as years of learning English in defining proficiency for learner corpus studies will be further discussed in the rest of this chapter.

Apart from large-scale projects of learner corpora mentioned above or small corpora built by individual researchers, one way to systematically collect learners' performance data is through the administration of language exams. For instance, Cambridge ESOL has not only constructed its own Cambridge Learner Corpus (CLC) from exam scripts[6] (Boyle & Booth, 2000) but also increasingly employed corpus-based approaches in versatile testing-related studies such as developing and validating the BEC (Business English Certificate) wordlist by investigating various corpora, CLC included (see Ball, 2001; Ball, 2002). In addition, a few studies have been conducted based on collections of exam scripts from Cambridge ESOL examinations (e.g. Banerjee et al., 2007; Kennedy & Thorp, 2007; Mayor et al., 2007). These learner corpora established upon examinee performance, however, are usually of a rather small size in comparison with native corpus research. Yet the results have cast light on writing ability as well as the distinguishing characteristics in different stages of interlanguage. Some of their findings will also be discussed in more detail in the following sections.

As Leech (1998, p. xvii) points out, learner corpora make it possible to investigate learner language from both negative and positive perspectives, i.e. 'what did the learner get

[6] The Cambridge Learner Corpus contains more than 135,000 exam scripts from 130 different L1s and 190 different countries, and these numbers are still on the increase (see http://www.cambridge.org/elt/corpus/learner_corpus.htm, visited on 29 September, 2009).

wrong?' and 'what did the learner get right?' With regard to what native norms to be compared against, a dilemma arises concerning whether we should choose native expert writing or native novice writing. According to Lorenz (1999), comparing learner language with native expert corpora is 'both unfair and descriptively inadequate' (p.14). On the other hand, if using native peer corpora as the native norm, i.e. texts produced by native students (native novice writing), it does not seem to be the case that native student language would be the optimal English standard if we would like to contrast the infelicity in learner language with proper native language. Unfortunately, it appears that a great number of 'Contrastive Interlanguage Analysis' (CIA) studies as termed by Granger (2002, pp. 12-13) made use of only native novice writing in comparison with learner writing (e.g. Aijmer, 2002; Altenberg & Tapper, 1998; Granger & Rayson, 1998; Granger & Tyson, 1996). This is probably due to the easy accessibility of LOCNESS (Louvain Corpus of Native English Essays), a corpus of native novice writing compiled in parallel with the ICLE corpus. As the result of difficulty in compiling a native reference corpus with the researchers' own resources, it is not surprising that very few studies have adopted native expert writing as the benchmark (e.g. Chen, 2006; Lorenz, 1999).

One problem accompanied by this issue is the determination of *overuse*, *underuse*, and *misuse*, which very often appear in CIA research to pinpoint the idiosyncrasies in learner language as to whether certain words or phrases occur more frequently (overuse) or less frequently (underuse) or whether their usage is erroneous (misuse) in comparison with native norms. However, there are problems associated with considering native student writing as the 'norm' as we generally do not expect British or American students to write perfect essays which can be used as a model even though English is their mother tongue. In this thesis, therefore, it is also decided to include two sets of native norms (both native expert writing and native peer writing) to be compared with learner writing so as to genuinely chart the

difference and/or similarity between native and non-native language use. The assumption underlying such comparative investigation is that learner idiosyncrasies can be identified when compared to both sets of native norms and hopefully can be rectified towards native-likeness when the idiosyncrasies are made explicit in language pedagogy.

After the native norms have been decided, this thesis also takes advantage of the notions of overuse and underuse in the comparison between groups of different proficiencies. As mentioned earlier, various learner corpus studies have drawn heavily on these notions in comparing native novice writing and learner writing. For example, Ringbom (1998) compared word frequencies between advanced learners' written English of seven different L1s in ICLE and native student writing in LOCNESS. The results show that all learner groups overused certain types of words such as auxiliaries, personal pronouns, and conjunctions, some core nouns (e.g. *time*, *way*, *people*, and *things*), some core verbs (e.g. *think*, *get*), and underused other types of words such as the demonstratives *this* and *these* and the prepositions *by* and *from*. The effect of the overuse and underuse, coupled with a limited range of vocabulary, not only dilutes the information content (*ibid*, p. 50) but also gives an impression of verbosity, dullness and repetition. Petch-Tyson (1998) also compared a native corpus of written American English with four L2 English learner groups of different L1 European languages from ICLE despite not explicitly using the terminology of overuse and underuse. For the purpose of comparing the extent of writer/reader visibility, she compared the frequencies of the defined features such as first and second person pronouns and fuzziness words (e.g. *kind/sort of, and so on)*. Petch-Tyson concluded that all the L2 groups of learners tended to express stronger interpersonal involvement, regardless of their L1, than the American writers investigated. As can be seen, due to the constraints of accessible data, many of the learner corpus studies have been confined to 'advanced' learner language compared with native language. One of the exceptions is Lorenz's study (1998), in which the researcher

used age as the variable to control linguistic proficiency. Two subcorpora of learner essays, one contributed by German teenagers (aged 16-18) and the other by German university students of English (aged 20-25), were compared against two other subcorpora of native essays, again, one produced by native British teenagers (aged 15-18) and the other by native British undergraduates (aged 19-23). The findings reveal that overall German learners of English demonstrated the peculiarity of excessive adjective intensification in comparison with British students, which leads to a style of unnatural overstatement. Furthermore, this 'over-zealousness' to impress the readers by employing undue intensifying adjectives does not seem to decrease with the development of learner proficiency.

Compared with the numerous studies which investigated L2 English of L1 European languages (probably as a result of the easy accessibility of ICLE), the number of studies based on L2 English corpora produced by L1 Chinese speakers is relatively smaller, and most of them used the learner data from Hong Kong students.[7] Take Hyland & Milton's study on qualification and certainty in L1 and L2 students' writing (1997) for example. By comparing the occurrences of epistemic lexical items (words only, no multi-word units included) in native and non-native corpora, they found that L2 English learners from Hong Kong demonstrated a more limited range of epistemic items and also made stronger commitment to the claims and propositions. Flowerdew (2000) looked at a subcorpus from the HKUST (Hong Kong University of Science and Technology) learner corpus of written data. She pointed out two types of common problems in the L2 English writing produced by L1 Cantonese-speaking students in Hong Kong. One is referential errors such as inappropriate collocation or multi-word units (termed *polywords* by Flowerdew). The collocations involved with delexical verbs such as *do*, *make*, and *have* were found to be particularly problematic.

[7] The ICLE subcorpus of L1 Chinese learners also comes from a number of universities in Hong Kong.

Meanwhile, when there are formulaic sequences such as *by means of* available, some students used a less formulaic paraphrase such as *by using* or *with the aid of*. The other problem with L2 English writing identified by Flowerdew is pragmatic infelicities, relating to learners being overtly direct and unhedged, a finding which also corresponds to Hyland & Milton's study (1997). Other studies which analysed the use of connectors in L2 writing in Hong Kong generally reported the overuse of connectors (e.g. *on the other hand*, *so*, *thus*) (Bolton et al, 2002; Field & Yip, 1992; Milton & Tsang, 1993). Although these studies already illustrate some interesting findings, it would be a misconception to consider L2 English learners from Hong Kong to sufficiently represent L1 Chinese students. On the one hand, Cantonese is one branch of the Chinese language family, just as the official Mandarin Chinese spoken in both China and Taiwan. Because Cantonese, as well as Taiwanese, can be rather mutually unintelligible with other varieties of Chinese, it is not clear whether the impact derived from different L1 Chinese language varieties would be the same, or similar, on L2 English learning.[8] On the other hand, the fact that Hong Kong used to be a British colony might have had a much greater impact on English learning than the issue of L1 Chinese varieties, as the English language has been far more frequently used in Hong Kong than any other L1 Chinese spoken regions. To my knowledge, quite a few secondary schools and universities in Hong Kong use English as the only medium language in the classrooms. The assumption that learners in Hong Kong could be much more proficient than learners from other L1 Chinese regions such as China or Taiwan is confirmed by the IELTS candidate performance report in 2008/2009 (Cambridge ESOL, 2009), in which candidates from Hong Kong received a much

[8] Taiwanese, as well as Mandarin, is spoken in Taiwan. Similarly, different varieties of Chinese languages are still used in China despite Mandarin as the official language. Whether, or to what extent, these Chinese language branches should be regarded as a language in its own right, as opposed to a dialect, is a linguistically and culturally complicated issue, which is far beyond the scope of this thesis.

higher overall IELTS band score (6.31) than Chinese (5.46) or Taiwanese students (5.66). In the learner writing investigated in this thesis, as the data comes from two existing corpora, it is not easy to control the variable of learner backgrounds. However, from the ethnographical information of learner data selected, at least it appears that the distribution of learners' origins is not dominated by any L1 Chinese variety group. The learner data used in this thesis will be described in two analysis chapters respectively (Chapters 5 and 6).

The learner corpus research addressed here is but a brief overview. More learner corpus studies which dealt with phraseology will be further discussed in Chapter 3. In the next section, I summarise how the results from learner corpus research can be applied in language teaching and learning.

2.1.3 Corpus Research & Second/Foreign Language Learning

Aston, Bernardini, & Stewart (2004) distinguished three types of relationships between corpora and language learners: corpora BY learners, corpora FOR learners, and corpora WITH learners. Learner corpora discussed earlier obviously fall into the category of the first type of relationship. As James (1992) remarked, 'the really authentic texts for foreign language learning are not those produced by native speakers for native speakers, but those produced by learners themselves' (p. 190). The ICLE corpus is an exemplar of illustrating corpora of this kind. An unprecedented collaboration between learner corpus research and English Language Teaching (ELT) publication is a section on academic writing in the second edition of the *Macmillan English Dictionary for Advanced Learners* (Rundell, 2007). In the 30-page long section on 'Improving Your Writing Skills', the ICLE researcher team demonstrated how learner corpus analysis can contribute to language learning by informing dictionary users in respect to a number of rhetorical features which have been revealed from the contrastive analysis between native and non-native corpus data.

With regard to corpora FOR learners, they refer to corpora which are designed to

facilitate language teaching and learning by providing descriptions of the target language in a specified context that corresponds to learners' needs. The *TOEFL 2000 Spoken and Written Academic Language* (T2K-SWAL) corpus or the *Michigan Corpus of Academic Spoken English* (MICASE)[9], for example, provide university-setting data which can well inform the syllabus design of EAP programmes or language tests. The former, T2K-SWAL, is a 2.7-million-word academic corpus which contains ten different spoken and written registers that students encounter most often in the settings of American universities. Biber and his colleagues have explored the linguistic variations in these different university contexts, ranging from lexical to syntactic aspects (Biber, 2006; Biber et al., 2004). One study drawn from the T2K-SWAL corpus data reports the differentiating features of lexical bundles in university teaching and textbooks (Biber, Conrad, & Cortes, 2004), which will be further discussed in Chapter 3. With a similar design, MICASE is a 1.8-million-word spoken corpus which records various academic speech events taking place at the University of Michigan, including classroom events such as lectures or lab sections and non-class events such as colloquia or meetings. Some informal instructional materials, both ESL/EAP teaching and ESL self-study materials, have been developed on the basis of MICASE data and are available online (http://lw.lsa.umich.edu/eli/micase/teaching.htm). Another example is a corpus composed of EFL textbooks compiled by Romer (2004), which functions as a 'pedagogical corpus' as defined by Hunston (2002, p. 16). The researcher worked on one case

---

[9] MICASE can be accessed via http://quod.lib.umich.edu/m/micase/ (visited on 25 June, 2009). Another corpus, the *Michigan Corpus of Upper-level Student Papers* (MICUSP), is also being constructed at the University of Michigan (http://lw.lsa.umich.edu/eli/eli1/micusp/index.htm, visited on 5 August, 2009). MICUSP aims to collect students' writing samples from both undergraduate and graduate levels and also from native and non-native speakers across the university, which does not seem to fit into either the profiles of corpora BY learners or corpora FOR learners.

study of *if*-clauses because the use of tense forms in the conditional constructions is often considered to be difficult for L2 learners. The result indicates some striking over- and under-representation of certain tense form sequences between textbook English and real-life English extracted from BNC data (the British National Corpus), which suggests that there is still much room for improvement in terms of authenticity in teaching materials.

The third type of relationship, corpora WITH language learners, is concerned with the activities designed on the basis of using corpora for language teaching and learning. This kind of corpus application can be exemplified by data-driven learning, as explained in detail by Hunston (2002, pp. 170-177). Basically corpora are used as a learning tool in this sense, and learners are often encouraged to discover the patterns emerging from corpus data themselves rather than being explicitly told about the properties of linguistic features. Recently researchers and practitioners have also been gradually interested in how phraseology can benefit ESL/EAP learning, but explicit learning is usually adopted as opposed to implicit learning. Two studies reporting the teaching of formulaic sequences in the classroom will be illustrated here. Based on frequency through a corpus search, Haywood & Jones (2004) selected approximately 80 formulaic phrases from EAP course books and applied them to a variety of activities in an L2 EAP classroom. Cortes (2006) experimented with 35 frequency-based formulaic sequences at a writing intensive history class of native American students. Both studies reported that the explicit instruction, unfortunately, did not make the students significantly increase the use of taught formulaic sequences in their writing, although the awareness of such formulaic sequences appeared to be raised over the period of instruction, which was as short as 10 weeks in both studies. It has to be noted that phraseology-teaching studies of this type have just received more attention recently as phraseology itself has just become a subject of research in its own right for the past decades. The impact of teaching formulaic sequences on language acquisition, explicitly or not, still

requires more empirical underpinnings. Discussion with regard to computer-derived phraseology in language learning will be further addressed in Chapter 7.

As can be seen, corpus research can make a great contribution to language learning from various perspectives. Below we will turn to second language research and language testing studies and discuss why the notion of learner proficiency should be incorporated into learner corpus research.

2.2 Developmental Studies and Proficiency Levels

In the domain of second language research, four approaches have been identified in describing learner language (Ellis, 1994):

- the study of learners' errors;
- the study of developmental patterns;
- the study of variability; and
- the study of pragmatic features.

Within the research which aims to search for developmental patterns in interlanguage, there are two types of developmental studies: developmental sequence studies and developmental index studies (as discussed in Wolfe-Quintero, et al., 1998, pp. 2-3). The former investigates the acquisition orders of morphosyntactic features while the latter generally observes learners' language development in terms of fluency, accuracy and complexity. The current study could be classed as the latter type, looking at learners' language development by addressing the discourse aspect of learner language, which has been relatively ignored in previous studies.

In past decades, there have been a large number of studies working on ESL/EFL writing development. Wolfe-Quintero et al. (*ibid*) compared 39 studies on second language development in writing and over 100 measures which gauged the development of learners at

known proficiency levels. These traditional second language studies, nonetheless, are largely small-scale in comparison with learner corpus research. Some studies investigated as few as 16 students' writing (e.g. Casanave, 1994), and interestingly only the number of students contributing to the studies were reported as opposed to the text size (word count). Although the meta-analysis conducted by Wolfe-Quintero et al. did not cater for aspects of discourse in learner English targeted by this thesis, it revealed some valuable observations about this type of developmental index study. For example, Wolfe-Quintero et al. pointed out that one of the principal goals of developmental index studies was to 'compare developmental measures with an independent measure of language or writing proficiency by means of correlations, t-tests, or analysis of variance' (*ibid*, p. 6). In other words, how proficiency levels were determined in the studies plays a key role in the discriminative power of developmental measures, and this is also the primary issue which will be discussed in this section.

As discussed in Section 2.1.2, there are variables concerning compilation of a learner corpus. Among those variables, learner proficiency is one major issue that is not well controlled in most learner corpus studies, and the condition in traditional SLA research, unfortunately, is also similar. In practice, proficiency is generally conceptualised through various ways such as rating scales, standardized tests, programme levels, school levels or classroom grades. While Quintero et al. (*ibid*, p.9) claimed that programme levels might be the most valid method to determine proficiency levels, here it is argued that marking EFL/ESL writing samples with a well-recognised rating scale such as the Common European Framework of Reference (CEFR) would be more sensible than any other methods that define learners' proficiency levels based on extra-linguistic judgments. This is because the researchers investigating second language writing development use merely one sample of writing from individual learners as the source of data for analysis and nothing else. Needless to say, language learning is never a linear phenomenon in every aspect of language. For

example, a learner might be proficient in every aspect of a target language but writing. If we determine his or her proficiency with a set of programme levels which concerns not just writing but also reading, listening and speaking, it is very likely that with the overall proficiency level this learner is assigned to, the information about his or her individual writing ability will be sacrificed. Accordingly, any linguistic analysis conducted on the basis of the overall proficiency would have been established upon a possibly incorrect assumption about the subject's writing ability, thereby undermining the validity of analysis. In addition, people learn at different rates and have different abilities so there can be a great deal of variation within one single year. An external criterion such as years of learning or programme levels, therefore, entails too much potential variability unknown to researchers. If the determination of proficiency levels is not reliable and specific to the linguistic aspect to be investigated, then the results of analyses will run the risk of losing their validity.

Since a rating scale would be employed to determine the proficiency levels of interlanguage for the current research, we need to comprehend better what lies behind the conception of 'scale'. In actuality, scales have been commonly used in rating and reporting language test performance. Alderson (1990, pp. 72-74) indicated three purposes of scales of language proficiency:

(1) *user-oriented*: reporting;

(2) *assessor-oriented*: guiding the rating process; and

(3) *constructor-oriented*: guiding the construction of tests at appropriate levels.

One key point in the reporting function of scales is 'to cover the notion that test results are typically ranges of possible scores, rather than precisely defined, invariable performances' (*ibid*, p. 73).

Indeed, when we deal with the data from a learner corpus in which each piece of candidates' performance is given a level, the notion of test results as 'ranges' instead of

'precisely defined' points is particularly important in the sense that one of the goals in the aforementioned developmental studies is to distinguish learner performance at different bands or levels. In other words, a script awarded Band Six in IELTS (International English Language Testing System) might be at the borderline between seven and six, i.e. a strong six, at the borderline of six and five, i.e. a weak six, or it could be merely a typical six. However, it is assumed that the scripts in Band Six must be somewhat distinct from those in Band Seven and Band Five, and the distinction is what the researchers strive to find out. In theory, if the data size is large enough, then it should be easier to spot the distinctions between bands, and this is why corpus-based approaches, which are used to analyse large amounts of naturally-occurring data, are useful to this type of research. Most of the currently available rating scales, to my knowledge, are established upon practitioners' intuitive perceptions rather than empirical evidence supported by learners' language development (see North & Schneider, 1998); hence, the research aiming for providing such empirical evidence from learner data would offer valuable information for devising or validating a scale of language proficiency.

2.3 Learner Proficiency in SLA, Learner Corpus Research, and Language Testing

It has been increasingly recognised that the notion of proficiency level has not been well-constructed in second language research. Thomas in a review article (1994) compared 157 studies with respect to assessment of L2 proficiency and concluded that sometimes the target language proficiency is poorly controlled to the extent that 'it limits the generalisability of research results.' After more than a decade, however, there has been little progress in this regard in second language research, including the newly developed sphere, learner corpus studies. Researchers still generally resort to extra-linguistic judgments in determining

proficiency, e.g. programme levels, which can easily undermine any empirical claims. Take the International Corpus of Learner English (ICLE), one of the most renowned learner corpora, for example. The learner essays with claimed 'advanced' proficiency in ICLE are broadly defined as essays written by the university students of English in their 3rd or 4th year of study. Making use of the ICLE data, Dagneaux et al. (1998) determined the proficiency levels of two continuous development stages, intermediate and advanced, with the curriculum separated by a 2-year gap, which is considered here to be too general and thus unreliable. With the aim of identifying learner levels with ICLE data, Pendar and Chapelle (2008) used the variable 'years of studying English' instead to divide ICLE essays into three proficiency groups. In search of developmental indicators, they then compared the three groups with various measures frequently utilised in SLA studies, such as type/token ratios or mean sentence length. Nonetheless, it was found that the reliability of proficiency defined this way was not sufficient, which unfortunately could undermine the validity of analysis results to a very large extent. They concluded that the corpus 'proved to be very difficult to use for this purpose because of the lack of information it contained about the evaluation of the essays' (*ibid*, p. 204).

In the past few years, there has been increasing research focus on the language features that distinguish learners' performance across proficiency levels, resulting in collaboration between researchers from the fields of language testing and SLA. The studies with empirical data retrieved from candidate scripts in high-stakes exams such as IELTS more or less include discourse features such as coherence and cohesion in the investigation of learner language development (see Table 2.1).

Table 2-1 Previous studies based on learner corpora of exam scripts from Cambridge ESOL (cited from Banerjee et al. (2007) with modification)

Study	No. of scripts	Corpus size (words)	Band levels of exams investigated (No. of scripts)	Task(s)	Versions of test	L1s
Banerjee et al. (2007)	550	132, 618	IELTS 3 to 8	Tasks 1 and 2	26	Chinese and Spanish
Hawkey & Barker (2004)	288	53,000	FCE (108), CAE (113) and CPE (67)	Guided Writing	1 (FCE)	Not reported; presumably mixed
Mayor et al. (2007)	186	56,154	IELTS 5 (100) vs 7/8 (86)	Task 2	2	Chinese and Greek
Kennedy & Thorp (2007)	130	35,464	IELTS 4/6 (50/50) vs. 8/9 (18/12)	Task 2	1	Reported as unknown; presumably mixed

For instance, Banerjee et al. (2007) aimed to identify the defining linguistic characteristics with regard to *cohesive devices used*, *vocabulary richness*, *syntactic complexity*, and *grammatical accuracy*. Their findings suggested that all the above criteria except for the syntactic complexity measures investigated are 'informative of increasing proficiency levels'. With the aim of developing a common scale for the assessment of writing in Cambridge Main Suite, Hawkey (2001) briefly introduced the rationale and methodology in a project which attempted to explore the definitive features in the writing performance of Cambridge ESOL users across FCE, CAE and CPE.[10] An initial analysis of the examinee scripts suggests that the written performance of these three levels can be distinguished by the impact on the readers which derives from several features, including *vocabulary*, *collocation*, *idiom use*, *pace variation*, *organisation structures*, and among others. In a follow-up paper, Hawkey and Barker (2004) described in detail how they adopted intuitive methods, qualitative methods, and quantitative methods proposed by CEFR and grouped their findings

[10] FCE: First Certificate in English; CAE: Certificate in Advanced English; and CPE: Certificate of Proficiency in English.

into versatile distinguishing features. The features explored included *impact*, *fluency*, *organisation*, *lexico-grammatical accuracy*, *sophistication of language*, *the lengths of whole script*, *sentence and paragraph*, *title use*, *vocabulary range*, and *words in concordance and collocations*. Among studies of this type, Kennedy and Thorp's project (2007) is probably the one that has considered aspects of discourse most thoroughly in learner language. Working with IELTS examinees' performance in Task 2 (argumentative essay writing) across several band scores, the researchers looked at a variety of features such as *rhetorical questions*, *modality items*, *discourse markers*, *subordinators* and *coordinators*, *boosters* and *downtoners*. One of their major findings is that compared with the examinees who received lower band scores, the more proficient IELTS writers used lexico-grammatical markers (e.g. *however*), enumerative makers (e.g. *firstly*), and subordinators (e.g. *because*) less frequently, which appeared closer to native-speaker use in this respect. In addition, the advanced IELTS writers were found to exhibit a wider range of lexis including idiomatic language (e.g. *guard against*, *whole host of opportunities*, *if and only if*), which was nearly absent in the scripts from the lower levels, although the researchers did not specify how they defined 'idiomatic language'. Despite the small number of samples investigated (130 scripts containing 35,464 words in total), their findings underpinned the argument that there is some sort of linear relationship underlying the discourse aspect in learner language development.

2.4 Conclusion

In this chapter, as we have seen, there has been some substantial progress in second language development studies and learner corpus research. Yet the notion of learner proficiency determination does not seem to be in good control in the areas aside from language testing. Most L2 corpus-based studies have generally compared 'advanced' learners' written performance with native novice writing, and the determination of 'advanced' level, as discussed earlier, is built upon assumptions which require further empirical evidence – e.g.

the longer learners study English, the more proficient they would be. This assumption takes into account neither individual learner difference nor the quality of performance produced by individual learners, which would certainly have a great impact on the results of analyses. Moreover, the fact that the majority of learner corpus research covers only a small section of learner proficiency (i.e. 'advanced' level) apparently cannot provide much information with regard to the lengthy process of second language development. It would be of great benefit to SLA as well as language testing research if the new corpus approaches can also be applied to describe learner language across various development stages.

As for the few developmental studies which have investigated learner writing performance with authentic examinee scripts, the corpus size is as a whole rather small in comparison of 100,000-word ICLE components contributed by individual L1 populations, particularly when the corpus size shown in Table 2.1 is broken down into individual subcorpora representing different proficiency levels. Setting out from the developmental perspective and determined to define learners' proficiency levels based on linguistic performance, this thesis therefore hopes to bridge the gap by means of large quantities of learner data and corpus approaches to investigate the discourse aspect of learner writing across proficiency levels. The corpus-driven and corpus-based approaches adopted will be introduced in the next two chapters.

Chapter 3 Theoretical and Operational Framework

This chapter will start with an introduction of phraseology and the status of *lexical bundles* as a frequency-based approach to extracting recurrent word strings. Following the introduction of phraseology is an account of the criteria which have been used to determine a lexical bundle in previous studies, and these criteria will also form the basis for the current study. Then the corpora investigated in this project will be briefly introduced because word combinations retrieved from these corpora will repeatedly emerge as the examples illustrated in the extraction procedure described in the rest of the chapter, which includes both automated extraction and manual examination. As a whole, this chapter aims to address the first methodological/procedural research question (see section 1.3) by searching for the optimum thresholds and procedures in terms of corpus size and frequency of bundles when investigating lexical bundle usage for the corpora investigated in this thesis.

3.1 The Study of Phraseology: Word Co-occurrence

In recent decades, researchers have become increasingly interested in how words co-occur in discourse to form formulaic units (e.g. *last but not least*, *things like that*, *pay attention*, *as well as*, etc.). Some linguists start from a theoretical perspective, claiming that multi-word combinations are stored as particular patterns in our mental lexicon while others build their arguments on the basis of empirical data, adding weight to the significance of multi-word expressions in language acquisition. For instance, Altenberg (1998) in his exploration of the London-Lund Corpus estimated that 80% of the words in the corpus formed part of a recurrent word combination. As Wray (2002, p. 9) observes, however, there has been 'the problem of terminology' to describe the phenomenon of word co-occurrence. As a matter of fact, it is never easy to reach a general conclusion from past studies in this area on the ground that the different perspectives and approaches having been applied make it extremely difficult

to compare the findings of numerous studies that have worked on the seemingly similar topics yet with various terms. On the one hand, various terms are used to refer to similar or even the same notion of word co-occurrence. On the other hand, the same term might be used in different ways by different scholars. Some examples of such terms include *clusters* (used in the corpus tool *WordSmith* and many other studies such as Hyland, 2008a; Mahlberg, 2007; Schmitt et al. 2004), *recurrent word combinations* (Altenberg, 1998; De Cock, 1998), *phraseology* (Howarth, 1998a, 1998b), *collocation* (Gledhill, 2000; Granger, 1998b; Luzon Marco, 2000; Nesselhauf, 2003, 2005), *phrasicon* (De Cock et al., 1998), *prefabricated patterns* (Granger, 1998b), *chains* (Stubbs, 2002), *formulaic language* (Oakey, 2002), *lexical phrases* (Li & Schmitt, 2009), *n-grams* (Fletcher, 2003-2006; Stubbs, 2007a, 2007b), and *lexical bundles* (e.g. Biber & Barbieri, 2007; Biber & Conrad, 1999; Cortes, 2002; Hyland, 2008b). As addressed in Chapter 1, however, 'phraseology' (Cowie, 1998b; Granger & Meunier, 2008; Meunier & Granger, 2007) and 'formulaic sequences/language' (Schmitt, 2004; Wray, 2002, 2008) are two umbrella terms probably used most often for various types of word associations. For a detailed discussion regarding the fuzzy terminology and typologies of word co-occurrence, please see Wray (2002) and Granger & Paquot (2008). Overall speaking, the study of how words co-occur in language is generally regarded as a complex and sophisticated domain which requires interdisciplinary collaboration. For example, Gries (2008) pointed out that phraseologism actually overlaps to a very large extent with other linguistic frameworks such as Cognitive Grammar or Construction Grammar.

Despite the disputable issue of defining different kinds of word co-occurrence and varying methodologies, word co-occurrence basically pertains to how the choice of one word can affect 'the choice of others in its vicinity' – the idiom principle addressed by Sinclair (1991, pp. 110-115). Traditionally, phraseology is considered to be a continuum of word combinations with various degrees of fixedness, ranging from the most fixed and opaque pure

idioms (e.g. *blow the gaff*), figurative idioms (e.g. *blow your own trumpet*), restricted collocations (e.g. *blow a fuse*), to the most variable and transparent free combinations (e.g. *blow a trumpet*) (see Cowie, 1981, 1998a). Generally speaking, the most idiomatic word combinations (or the most non-compositional ones) such as *blow the gaff*, which cannot be understood or analysed from the individual components (i.e. *blow*, *the*, and *gaff*), are those that have received more attention in conventional phraseology studies. The traditionally defined phraseological language is usually researched with a pre-existing repertoire of phraseological units to compare against the corpus data (e.g. Moon, 1998b) or to adopt native-speakers' intuition/preference to demarcate the phraseological sequences in the texts (e.g. Erman & Warren, 2000; Li & Schmitt, 2009).

With the rapid development of corpus linguistics in the past few decades, researchers have begun to take quantitative approaches in examining how words co-occur to form prefabricated units, resulting in systematic and quantitative analyses of various forms of formulaic sequences. Stubbs (2002) distinguished two quantitative methods of studying phraseology in English: one involves the determination of collocations with certain extraction statistical measures (e.g. mutual information, log-likelihood, t-test), and the other deals with continuous word combinations retrieved with specified frequency and distribution criteria (e.g. occurring at least 40 times per million words in five texts or more in a particular corpus). This second method, which investigates continuous word combinations, has been extensively employed in corpus studies under various terms mentioned above, *clusters, phrasicon, chains, n-grams, recurrent word combinations*, and *lexical bundles*. The recurrent sequences retrieved with specified frequency and dispersion thresholds are fixed multi-word units (e.g. *in the case of, it is possible to*) which are found to have customary pragmatic and/or discourse functions, used and recognised by the speakers of a language within certain contexts. Meanwhile, these highly frequent sequences largely stretch across the borderline between

lexis and syntax, functioning as lexico-grammatical building blocks of discourse. This methodology is considered to be a frequency-based approach of determining phraseology (see Granger & Paquot, 2008).

From a psycholinguistic viewpoint, formulaic language has been found to have 'a processing advantage over creatively generated language' for non-native as well as native speakers (Conklin & Schmitt, 2008, p. 72), although different psycholinguistic studies have used various types of formulaic language such as idioms (e.g. *take the bull by the horns*) or non-idiomatic phrases (e.g. *as soon as*) as the target forms. A particularly inspirational study was conducted by Jiang & Nekrasova (2007), in which they utilised corpus-derived recurrent word combinations as materials in two online grammaticality judgement experiments. Their findings provided 'prevailing evidence in support of the holistic nature of formula representation and processing in second language speakers' (*ibid*, p. 433). Schmitt et al. (2004) investigated the psycholinguistic validity of corpus-derived recurrent clusters as well, but they concluded that not all the clusters are psycholinguistically valid although their results shared some similarities with Jiang & Nekrasova (2007).

Taking into account the overall supporting psycholinguistic evidence (despite not being conclusive) and the large size of linguistic data, the current study has therefore decided to adopt the frequency-based approach, i.e. to explore the data through defining and examining recurrent word sequences with the corpus tool *WordSmith 4.0* (Scott, 2007). *Lexical bundles* will be used as the primary term throughout the whole thesis as it is the term used by Biber and his colleagues in a series of studies which the theoretical and analytical framework of the current study is established upon. Another term *recurrent word combinations* may henceforth be used interchangeably with *lexical bundles* in this thesis because its literal meaning is transparent and also because it has been used with the same sense as *lexical bundles* by a couple of other researchers (Altenberg, 1998; De Cock, 1998).

3.2 Lexical Bundles in L1 and L2 Studies

In general, the automated approach of identifying lexical bundles appears to have been initially applied to spoken language (e.g. Altenberg, 1998) and widely utilised by De Cock in comparing L2 speech with native speech (De Cock. 1998, 2004, 2007; De Cock et al., 1998), although other terms such as *phrasicon*, *recurrent word combinations*, or *preferred sequences of words* were used in these studies. Furthermore, De Cock took advantage of this approach in comparing L2 with L1 not only in speech but also in writing (De Cock, 2000). These comparative studies as well as Altenberg's pioneering research (1998), however, investigated repetitive chunks from an overall perspective of looking at all the continuous word combinations ranging from two-word to five-word lengths without further categorisation or refinement. The problem is that, as De Cock herself pointed out, not all the automatically extracted word sequences are qualified as prefabricated expressions because 'a structural classification and a thorough functional investigation of the combinations in context is required before they can be labelled as such' (De Cock, 2000, p. 59).

Working with an online interface *Phrases in English* (PIE) (Fletcher, 2003-2006), which offers computed information on recurrent phraseology retrieved from the British National Corpus (BNC), Stubbs (2007a, 2007b) set off from a more theoretical perspective to look into the nature of lexical bundles, which he termed *n-grams* following the terminology in PIE. Stubbs also illustrated two more functions provided by PIE, *p-frame*, and *PoS-gram* as well as *n-gram*. He explained the concepts of these terms as below (2007b, p. 166):

n-gram : a recurrent uninterrupted string of orthographic word-forms
p(hrase)-frame : a recurrent n-gram with one variable lexical slot
PoS-gram : a recurrent string of part of speech tags

Stubbs indicated that the most frequent 4-grams in the 100-million-word BNC (written data of 90 million words and spoken data of 10 million words) are prepositional

phrase fragments (e.g. *in the middle of*, *as a result of*) and four of the most frequent 5-PoS-grams are either parts of nominal phrases (e.g. *the end of the year*, *the other side of the*) or prepositional phrases (e.g. *at the end of*, *in the case of the*). Yet Stubbs's contribution in this area lies in his critical inspection of the status of recurrent phraseology. He challenged the traditional view of defining 'word frequency' as many words have a high frequency actually because they are part of many highly frequent phrases. He further pointed out that this frequency approach of retrieving repetitive strings of word-forms raises 'the classic problem of how to identify linguistic units', which is worth careful consideration when we attempt to apply the findings from the quantitative analysis in lexical bundles in the area of ESL/EFL teaching and learning.

In the lexical bundle studies conducted by Biber and his colleagues (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber et al., 2003; Biber et al., 2004; Biber et al., 1999), it has been found that conversation and academic prose present distinctive patterns of bundle distribution. For example, most bundles in conversation are clausal (verbal bundles) whereas academic prose is characterised with phrasal bundles (nominal and prepositional bundles).[11] In terms of discourse functions, the bundles found in speech are primarily used as stance expressions (to express modality or attitude such as *I don't think so*) or interactional markers (which orient to the listener such as *you know what*). In contrast, the bundles found in writing are mostly referential bundles (which reference specific attributes such as *in the context of*) and discourse organisers (which introduce or clarify topics such as *as a result of*). The structural and functional taxonomies developed by Biber and his colleagues form the

---

[11] It has been mentioned in Section 1.3 that the terms 'phrasal bundles' and 'clausal bundles' used by Biber et al. can be ambiguous as there are verb phrases as well as noun and prepositional phrases. Yet the distinction they made denotes that 'phrasal bundles' refer to only noun and prepositional combinations whereas 'clausal bundles' contain a verb component and possibly a pronoun.

analytical framework of the present thesis, which will be addressed in the next chapter.

Other studies of bundles have mostly focused on comparisons between expert and non-expert writing. Cortes (2002) investigated bundles in native freshman compositions and found that the bundles used by these novice writers were functionally different from those in published academic prose. In another study, Cortes (2004) compared native student writing with writing in academic journals, concluding that generally students did not use the lexical bundles identified in the corpus of published writing. Even if they did, students used these bundles in a different manner. Working with academic writing only, Hyland (2008b) indicated that there is disciplinary variation in the use of lexical bundles. He also investigated the role of lexical bundles in differentiating published and postgraduate writing and found that postgraduate students tended to employ more formulaic expressions than native academicians to display their competence (Hyland, 2008a). We will further compare the procedures and results between the current study and Hyland's (2008a) in Chapter 7. The corpus data and operational definitions of recurrent sequences investigated in the above studies will be summarised in the next section.

So far few studies of L2 written data have included structural and functional categorisations of lexical bundles. Although Hyland in his two studies (2008a, 2008b) included masters' theses and doctoral dissertations produced by L2 English students in Hong Kong, he did not set off from the perspective of second language learning. Instead, he treated L2 postgraduate writing as 'highly proficient' prose on the ground that the data in his corpus texts had been all awarded high passes. Different from Hyland's studies, the present thesis hopes to reveal the potential problems in second language learners' written performance from the viewpoint of frequency-defined phraseology by making comparison with native writing.

3.3 Determination of Lexical Bundles

According to Altenberg (1998), *recurrent word combinations* (a.k.a. *lexical bundles*) refer to linguistic patterns that meet the following criteria:

1. they have identical forms;
2. they are a continuous string of words; and
3. they occur more than once in the text or texts under examination.

Although thus far only a limited number of studies have worked on the notion of lexical bundles (see a comprehensive review of past studies in Table 3.1 and Table 3.2), several key criteria have been pinpointed with regard to how to generate a list of lexical bundles through automated corpus tools. The first criterion is the cut-off frequency, which determines the number of lexical bundles to be included in the analyses. As shown in Table 3.1 and Table 3-2, the normalised frequency threshold for large written corpora generally ranges from 20 to 40 times per million words (e.g. Biber et al., 2004; Hyland, 2008b) while for the relatively small spoken corpora, the raw cut-off frequency is often used, ranging from two to ten times (e.g. Altenberg, 1998; De Cock, 1998). The second issue concerns the length of word combination, usually 2-, 3-, 4-, 5-, or 6-word units. Four-word sequences are found to be the most researched length of bundles for writing studies, probably because the number of four-word bundles generally falls within around 100, which is of a manageable size for manual categorisation and concordance checks. The last criterion is the requirement that the combinations have to be repeated in different texts, usually in at least three or five texts (e.g. Biber & Barbieri, 2007; Cortes, 2004) or 10% of texts (e.g. Hyland, 2008a), which helps to avoid idiosyncrasies from individual writers/speakers.

As can be seen, frequency and dispersion thresholds adopted have varied from study to study. Even the sizes of corpora and subcorpora compared to each other have differed drastically, ranging from a huge corpus of over five million words to a small subcorpus of

nearly 40,000 words. Despite the discrepancy between studies that has been observed, such an overview lends itself for the present study to determining the scope for investigation. Four-word combinations, therefore, would be the target length of lexical bundles for the current project so that the results can be compared with those in the L1 studies. In addition, after repeated experiments with the corpus data in use, the cut-off frequency and distribution for locating the word sequences for investigation was decided to be four times or more (around 25 and 45 times per million words on account of subcorpora with different sizes investigated) occurring in at least three texts. Word combinations retrieved with these designated criteria were initially examined and found to be appropriate for this project in terms of their quantity and quality. The decision on such thresholds will be further discussed in Section 3.5.1 after a brief introduction of the corpora investigated in Section 3.4. A description of how the word combinations were automatically generated and manually filtered will also be given in Sections 3.5 and 3.6.

Table 3-1 Past studies on lexical bundles in written or written and spoken registers

Study	Corpus size (tokens) & Registers	L1/L2	Bundle length, frequency & dispersion threshold	No of Target Bundles Retrieved
Hyland (2008a)	Total: 3,455,000 **Written** • Published research articles: 730,000 (120 texts) • PhD dissertations: 1,900,000 (80 texts) • MA/MSc theses: 825,000 (80 texts)	L1 + L2	4-word bundles occurring 20 times per million words in at least 10% of texts	• 130 different bundles Research articles: 71 PhD dissertations: 95 Master's theses: 149
Hyland (2008b)	Total: 3,400,400 **Written** • Electrical engineering: 632,500 • Biology: 794,000 • Business studies: 844,400 • Applied linguistics: 1,129,400	L1 + L2	4-word bundles occurring 20 times per million words in at least 10% of texts	• 240 different bundles Engineering: 213 Biology: 131 Business Studies: 144 Applied Linguistics: 131
Stubbs (2007a, 2007b)	100-million-word BNC, including **Written** (90 million words) and **Spoken** (10 million words)	L1	3 times for n-grams	(A more qualitative study)
Biber & Barbieri (2007)	**Spoken** • classroom teaching: 1,248,811 • classroom management: 39,255 • office hours: 50,400 • student groups: 141,100 • service encounters: 97,700 **Written** • Textbooks: 760,619 • Course management: 52,410 • Institutional writing: 151,500 **Academic prose** 5,330,000	L1	4-word bundles occurring 40 times per million words in multiple texts (varying from at least 3 to 5 texts with the size of subcorpora)	**Spoken** • classroom teaching: 80 • classroom management: 90 • office hours: 60 • student groups: 40 • service encounters: 100 **Written** • Textbooks: 30 • Course management: 130 • Institutional writing: 95 • Academic prose: 20
Biber et al. (2004)	• Classroom teaching: 1,248,000 words • Conversations: 4 millions British English+ 3 million American English • Textbooks: 760,000 words • Academic prose: 5.3 million	L1	4-word sequences occurring at least 40 times per million words in at least 5 different texts	• Conversation: 43 • Classroom teaching: 84 • Textbooks: 27 • Academic prose: 19

Study	Corpus size (tokens) & Registers	L1/L2	Bundle length, frequency & dispersion threshold	No of Target Bundles Retrieved
Cortes (2004)	**Published writing** • History: 966,187 • Biology: 1,026,344 **Student writing:** • History: 493,109 • Biology: 411,267	L1	4-word bundles occurring at least 20 times per million words in 5 or more texts	•History: 54 •Biology: 109
Biber et al. (2003)	5 million words for each of **written** register (academic prose) & **spoken** register (conversation)	L1	3-word, 4-word, 5-word and 6-word bundles at least 20 times per million words in at least 5 different texts	Not reported
Cortes (2002)	360,704 words of **written** data (freshman composition, including descriptions, rhetorical analyses, research proposals and research papers)	L1	4-word bundles occurring at least 20 times per million words in 5 or more texts	93
De Cock (2000)	**Spoken** L1: 117,417 words L2: LINDSEI, 90,300 words **Written** L1: LOCNESS, 106,112 words L2: ICLE, 100,575 words	L1 vs. L2	2-word at least 12 times 3-word at least 6 times 4-word at least 4 times 5-word at least 3 times 6-word at least 3 times (approximately 10% of bundle types are included for each length)	No exact numbers reported, only presented in graphs

Table 3-2 Past studies on lexical bundles in spoken registers only

Study	Corpus size (tokens) & Registers	L1/L2	Bundle length, frequency & dispersion threshold	No of Target Bundles Retrieved
De Cock (2004)	**Spoken** : L1: LOCNEC, 90,300 words L2: LINDSEI, 41,000 words	L1 vs. L2	2-word at least 12 times 3-word at least 6 times 4-word at least 4 times 5-word at least 3 times 6-word at least 3 times	No exact numbers reported, only presented in graphs
De Cock et al. (1998)	**Spoken**: L1: 80,448 words L2: 62,975 words	L1 vs. L2	2-word at least 9 times 3-word at least 4 times 4-word at least 3 times 5-word at least 2 times	**L1**: 588 in 2-word, 512 in 3-word, 119 in 4-word, 39 in 5-word **L2**: 606 in 2-word, 508 in 3-word, 149 in 4-word, 43 in 5-word
De Cock (1998)	**Spoken** : L1: 57,000 words L2: LINDSEI, 41,000 words	L1 vs. L2	2-word at least 10 times 3-word at least 5 times 4-word at least 4 times 5-word at least 3 times	No exact numbers reported, only presented in graphs
Altenberg (1998)	**Spoken**: London-Lund Corpus: 500,000 running words	L1	At least 3 words occurring at least 10 times in the corpus	470 in total (for 3- to 5-word combination)

3.4 Corpora Investigated

Three corpora were used for the present study: the Freiburg-Lancaster-Oslo/Bergen Corpus of British English (FLOB), the British Academic Written English (BAWE) corpus, and the Longman Learners' Corpus (LLC). These three corpora will be briefly introduced in this section. For the sake of comparability, only part of each corpus was selected for investigation.

The FLOB corpus is a one-million-word corpus of written published British English compiled in the early 1990s, and it contains fifteen text categories. For the current thesis, only the category of academic prose, FLOB-J, was used as the representative group of expert academic writing. The category FLOB-J is composed of eighty excerpts from published academic texts retrieved from journals or book sections in a wide range of disciplines.

With regard to the student academic writing, part of the BAWE corpus was utilised. The BAWE corpus was released in 2008, and it contains approximately 3,000 pieces of proficient assessed student writing from British universities, which amounts to 6.5 million words in total. Two subcorpora were carefully selected from the BAWE corpus: one is the BAWE-CH, which contains L2 English essays produced by L1 Chinese students, and the other – BAWE-EN is a comparable subcorpus contributed by peer L1 English students. FLOB-J, BAWE-CH, and BAWE-EN were all used in the first modular study, and the corpus size for each is around 150,000 words. The selection of corpus data along with its ethnographical and linguistic information will be further illustrated in Chapter 5.

The last corpus is the Longman Learners' Corpus (LLC), a commercial corpus of learner English. It is a large computerised collection of documents written by learners of L2 English, comprising mainly essays and exam scripts with general topics from language schools, teachers, students throughout the world between 1990 and 2002. Only the pieces written by L1 Chinese learners of L2 English with appropriate topics, mainly being argumentative or expository, were chosen as the candidate samples. After a standard rating

procedure, an even smaller proportion of the selected L2 English samples ranging within two proficiency levels (CEFR-B2 and CEFR-C1) were included in the investigation. Two learner subcorpora representing one of the two proficiency levels accordingly came into being, and they were used in the second modular study with a corpus size of approximately 88,000 words for each. The rating procedure and selection of corpus data will be described in detail in Chapter 6.

Further details about the rationale of separating these corpus data into two modular studies and how these subcorpora[12] were compared will be fully discussed in the second part of methodology description, Chapter 4 Analytical Framework. However, how the cut-off frequency and distribution thresholds were determined and how the word combinations were extracted and manually examined will be discussed in the rest of this chapter, illustrated with the lexical bundles retrieved from the above five groups of writing.

3.5 Automatic Retrieval

3.5.1 Determination of Frequency and Dispersion Thresholds

As mentioned in Section 3.3, the cut-off frequency and dispersion used to define a recurrent word combination was set to four times or more in at least three texts (around 25 times per million words in Modular Study 1 and 45 times per million words in Modular Study 2). This is a decision made from repeated experiments with various frequency and dispersion rates, first starting with the corpora used in Modular Study 1 and then applied in the corpora investigated in Modular Study 2. The specified threshold resulted in an 'optimum' number of bundles, more or less around 100, which is considered to be sufficiently representative of the

[12] Although some different groups of writing defined in this thesis come from the same corpus, they will henceforth be treated and termed as independent (sub)corpora representing a specified group of writing.

corpora being examined and also a suitable size for manual examination and concordance checks. The process of repeated experiments and the formulation of the rationale behind the decision of this threshold will be described below.

First, the relationship between the threshold and the number of recurrent word combinations generated is considered. As can be seen in Table 3-3, although the corpora in Modular Study 1 (BAWE-CH, BAWE-EN, and FLOB-J) are of a similar size, the word count is still slightly different, and the normalised frequency thereby varies to an extent. Yet with the raw cut-off point set at four times in at least three texts, the number of lexical bundles accordingly generated falls within a reasonably manageable size of around 100 raw instances (types as opposed to tokens).[13] If the cut-off point, however, is first set at 20 times per million words occurring in at least three texts, the converted raw frequencies would be 3.3, 3.1 and 2.9 respectively for the three corpora and therefore would need to be rounded up or rounded down to three times (see Table 3-4). As a result, the number of lexical bundles retrieved with the frequency threshold of three times almost doubles to around 200 raw instances. Considering the time required for manual examination and concordance checks, this number of bundles was felt to be too large for investigation. At the same time, after an initial inspection of the KWIC (Key Words in Context, the concordance lines), it was found that when using a cut-off frequency of three times, the clusters retrieved contain many more undesired context-dependent bundles, which will be explained in the next section, when compared to using four times as the cut-off frequency. Determining a threshold is thus a tug of war between the amount of information and the degree of precision/representativeness. The stricter the threshold, i.e. the higher cut-off frequency and dispersion, the fewer bundles

[13] The instances here refer to bundle types, not tokens. They are considered 'raw' instances because they have not been filtered by taking 'overlaps' and 'context independence' into consideration. The concepts of these two terms will be addressed in the next two sections.

will be generated, although they are guaranteed to be of high frequency and good quality for investigation. In contrast, the looser the threshold is, the wider range of bundles will be retrieved for investigation, yet some of them might not be appropriate for analysis, and it is also very likely that much more time would be required for the researcher to manually filter the data. A subtle balance between these two conflicting factors has to be kept in mind as we want to acquire as much information as possible while the availability of time and manpower also has to be considered.

Table 3-3 Threshold of raw frequency of four times occurring in at least three texts

	Corpus size (words)	Raw frequency threshold	Converted normalised frequency per million words	Raw bundle types
BAWE-CH	146,872	4	27.2	89
BAWE-EN	155,781	4	25.7	120
FLOB-J	164,742	4	24.3	118

Table 3-4 Threshold of normalised frequency of 20 times per million words occurring in at least three texts

	Corpus size (words)	Normalised frequency per million words	Converted raw frequency threshold / rounded frequency	Raw bundle types
BAWE-CH	146,872	20	2.9 / 3	171
BAWE-EN	155,781	20	3.1 / 3	224
FLOB-J	164,742	20	3.3 / 3	238

The second factor that has to be taken into account is to determine which frequency threshold to be reported, the normalised frequency or the raw one. One might argue that an identical standardised threshold, such as 20 or 40 times per million words, should be applied for each of the corpora investigated as generally reported in the literature. Yet the reason why the current study did not establish the cut-off frequency in the conventional way, i.e. setting the normalised frequency threshold first and then switching to the raw frequency one, is also a decision that was reached after repeated experiments. During my experiments, a

discrepancy between the normalised cut-off frequency and the converted raw frequency was found. When a set normalized rate is converted to raw frequencies corresponding to different corpus sizes, it substantially affects the number of generated word combinations when comparing corpora of various sizes. Biber and Barbieri (2007), for instance, compared a 5.3-million-word written corpus with a 40,000-word spoken corpus. With the cut-off standardised frequency set at 40 times per million words, this means that the converted raw frequency threshold for the large written corpus is as high as 212 whereas the converted raw frequency threshold for the small spoken subcorpus is relatively much lower at 1.6. Researchers working on lexical bundles generally did not make it clear as to how they dealt with a raw cut-off frequency which contained a decimal point, but it is reasonable to assume that the decimals are either rounded up or rounded down. Yet rounding up 1.6 to 2 results in a normalised rate of 50 times per million words, which is 10 times more than the originally reported frequency threshold of 40 times. Reporting only the standardised frequency criterion, therefore, could result in misleading claims because a standardised cut-off frequency would inevitably lose its expected impartiality after being converted into raw frequencies, particularly when the corpus size varies drastically and the converted raw frequency has to be rounded up or down. In this thesis, it would be argued that both the raw cut-off frequency and the corresponding normalised frequency should be reported in order to transparently reflect the threshold adopted. For the sake of comparison, if the frequency threshold is set at 25 times per million words for the Modular Study 1, the converted raw frequency for each corpus is 3.7, 3.9 and 4.1 times, which are all still rounded up or down to 4 times (cf. Table 3-5 and Table 3-6). The frequency threshold thus would be reported as word sequences occurring at least four times (raw cut-off frequency) or 25 times per million words (normalised frequency) in each of the corpora in Modular Study 1.

Table 3-5 Raw and converted normalised frequency thresholds adopted

Corpus	Set raw frequency threshold	Converted normalized frequency (per million words)
BAWE-CH	4	27.2
BAWE-EN	4	25.7
FLOB-J	4	24.3

Table 3-6 Normalised and converted raw frequency thresholds for comparison

Corpus	Set normalized frequency threshold (per million words)	Converted raw frequency
BAWE-CH	25	4.1
BAWE-EN	25	3.9
FLOB-J	25	3.7

It is possible that the above issues might be a consequence of using small corpora in this project. In theory, the impact originating from the discrepancy between normalised and raw frequencies would be far less in researching large corpora of a similar size of over one million words. Indeed, some doubts might arise about using such small corpora of only approximately 88,000-150,000 words to retrieve lexical bundles. However, corpus size and number of lexical bundles yielded with different frequency and dispersion rates is a conceptually complex issue. In the same study discussed earlier, Biber and Barbieri (2007) explained in great detail how they took advantage of dispersion requirement when comparing some very small subcorpora of only 40,000 words to 50,000 words in the spoken register with a corpus of academic prose containing millions of words. In fact, in a large corpus even the least common lexical bundles occur in multiple texts, at least in 20 texts in Biber and Barbieri's study (*ibid*, p.268), and hence the dispersion threshold does not seem necessary at all for large corpora in determining lexical bundles. On the contrary, applying a distributional requirement as well as a cut-off frequency to small corpora appears to be very effective in

terms of filtering out the word combinations that do not distribute broadly enough in the texts. In addition, while numbers of bundle tokens are linear with corpus size, numbers of bundle types are not (the rationale of distinguishing bundle types and tokens will be discussed in Chapter 4), as the range of formulaic units in our mental lexicon is supposed to be finite in comparison with the infinite combinations of creative units. It is thus possible to retrieve representative frequency-driven phraseological language from a small corpus. Adopting a dispersion requirement is considered to be effective enough in remedying for the potential risk in using small corpora that certain word combinations reach the cut-off frequency on account of some writers' idiosyncratic styles or other contextual factors, which is more unlikely to take place when using large corpora.

In order to consolidate the assumption that corpus size does not essentially affect the generation of lexical bundles, an experiment was carried out on the whole BAWE corpus, which comprises approximately 6.5 million words across 2,761 texts. The normalised frequency threshold for defining a lexical bundle was set at an occurrence of 20 times per million tokens, which was converted to the raw frequency of 130 times in the complete corpus of 6.5 million words. Originally the other condition to determine a lexical bundle is that any recurrent word combination has to appear in at least five texts. However, as just discussed, the results showed that none of the target bundles failed to meet this criterion of dispersion owing to the enormous amount of data. In fact, the least frequent expression in the extracted bundles *it can be argued* distributes across as many as 93 texts, a number far higher than at least 20 different texts as reported in Biber et al's corpus of five million words. Another result of this experiment is the number of lexical bundles generated. Although the normalised frequency threshold for the whole BAWE, 20 times per million words, was lower than the frequency of 25 times per million words in BAWE-CH, BAWE-EN, and FLOB-J, the whole BAWE corpus (coded as BAWE-all) yielded the smallest number of lexical

bundles (see Table 3-7). As far as what has been observed, it can be asserted here that small corpora can still be used to generate quality lexical bundles, if the selection criteria are established appropriately. As a matter of fact, small corpora generally provide a wider range of lexical bundles by applying a higher frequency threshold, as shown in the current case.

Table 3-7 Corpora of different sizes and the number of the retrieved lexical bundles

	Word count	Cut-off Frequency (per million words)	Dispersion threshold	Raw bundle types
BAWE-CH	146,872	25 times (four times as the raw freq)	in at least three texts	89
BAWE-EN	155,781			120
FLOB-J	164,742			118
BAWE-all	6,506,995	20 times (130 times as the raw freq)	in at least five texts	89

3.5.2 Results of Automated Retrieval

The corpus analysis tool *WordSmith 4.0* (Scott, 2007) was used for various types of automatic analyses such as word counts, concordancing, or lexical bundles (termed as *clusters* in *WordSmith*). The cut-off frequency and dispersion thresholds, as discussed in the previous section, are set at selecting the word combinations that occur four times or more in at least three texts.

As can be seen in Table 3-8, the subcorpora used in Modular Study 2 is nearly half of those used in Modular Study 1 in terms of size (about 150,000 words versus 88,000 words). Yet it is still decided to keep the same threshold in Modular Study 2 in the sense that the numbers of word combinations retrieved have reached the extent that it requires great efforts for manual examination as the number of raw bundle tokens, 936 and 996 respectively, means that nearly 1,000 concordance lines have to be checked for each of the subcorpora CEFR-B2 and CEFR-C1 in Modular Study 2. The reason why the numbers of tokens in

Modular Study 2 is substantially higher than those in Modular Study 1 will be addressed in the next section. The cut-off raw frequency of three times (about 35 times per million words) was also tried; however, an initial inspection suggested that far too many context-dependent bundles would be included. Given that this thesis intends to examine around 100 bundle types in each corpus investigated, the identical threshold is hence adopted in both of the two modular studies.

Table 3-8 Corpora used in two modular studies, the threshold, and the number of retrieved lexical bundles

Study	Corpus	Corpus size (words)	Cut-off raw freq	Cut-off normalised freq (per million words)	Dispersion requirement	Raw bundle types	Raw bundle tokens
Modular Study 1	BAWE-CH	146,872	4	25	at least 3 texts	90	554
	BAWE-EN	155,781	4	25		120	757
	FLOB-J	164,742	4	25		118	749
Modular Study 2	CEFR-B2	87,970	4	45	at least 3 texts	164	936
	CEFR-C1	87,828	4	45		169	996

3.6 Manual Examination

After the automatic retrieval of four-word clusters with *WordSmith*, two more procedures were carried out so as to filter out the 'noises' in the retrieved data. 'Noises' here refer to the word combinations which are not desired for investigation. They are the word combinations that recur either because of the context in which they are present in, such as being part of essay topics, or because of two or more retrieved word sequences which overlap. The details will be further explained below.

3.6.1 Context Independence

After automatic retrieval of 4-word clusters with the corpus tool *WordSmith 4.0* (Scott, 2007), word sequences that contained content words present in the essay questions (e.g. *financial*

and non financial) or any other context-dependent bundles, usually incorporating proper nouns (e.g. *in the UK and*, *the Second World War*), were manually excluded from the extracted bundle lists. These context-dependent bundles have to be removed as they are not the 'building blocks' which carry a distinct discourse function intended by the current thesis for analysis. The concordance lines of bundles were checked whenever in doubt. This procedure has also been described in De Cock (1998), De Cock et al. (1998) and Milton (1998) but not in a series of bundle studies conducted by Biber and his colleagues.[14] Instances like this in the FLOB-J are *in the UK and* and *the Second World War*, and no context-dependent bundle was found in the BAWE-EN. As an L2 corpus, BAWE-CH does not show too much difference in comparison with the native data as it only contains four context-dependent bundles: *in the United States*, *financial and non financial*, *of goods and services*, and *of the company as*. As can be seen, the two word sequences *of goods and services* and *of the company as* are not exactly proper nouns, but the concordance lines suggest that these bundles bear on the course modules students attended, i.e. business-related modules despite the contributors' various disciplinary backgrounds documented in the BAWE.[15] After the filtering process, the BAWE-CH remains the group with the smallest number of lexical bundles (see Table 3-9).

---

[14] It is very likely for a very large corpus to have fewer or no context-dependent bundles, e.g. the Longman Spoken and Written English Corpus used in Biber et al. (1999, 2003), because the required cut-off frequency, say 20 times per million words, would be multiplied to 100 times in a 5-million-word corpus and therefore decrease the likelihood of generating the word combinations dependent on the context.

[15] Students might attend a module that is not directly relevant to their disciplines. For example, one student who contributed one of the essays where the bundle *of the company as* was retrieved came from the Department of Engineering, but the essay under investigation was submitted for the module of Financial and Cost Management.

Table 3-9 Number of lexical bundles before and after filtering out context-dependent bundles in Modular Study 1 (types)

Corpus	Raw instances	Filtered instances
BAWE-CH	90	86
BAWE-EN	120	120
FLOB-J	118	116

The phenomenon of context (in)dependence with regard to the bundles retrieved in Modular study 2 demonstrates a completely different picture. A great number of word sequences can be filtered out even without concordance checks because a quick scan of the bundle list reveals that a substantial part of the word combinations are context-dependent. A careful examination of the concordance data confirmed this impression. The result showed that those context-dependent bundles constitute over half of the automatically retrieved word combinations in CEFR-B2 and almost 75% of the combinations in CEFR-C1 (see Table 3-10).

Table 3-10 Number of lexical bundles before and after filtering out context-dependence in Modular Study 2 (types)

Corpus	Raw instances	Filtered instances
CEFR-B2	164	84
CEFR-C1	169	42

A few examples of context-dependent bundles in the CEFR-C1 subcorpus are *second language acquisition in*, *the Hong Kong government* and *the use of computers*. Some examples found in the CEFR-B2 subcorpus are *to the crown court*, *Tsim Sha Tsui is* and *the countryside is more*. As in Modular Study 1, these word combinations recur in the light of the context they take place in. Sometimes the context-dependent bundles contain proper nouns related to the socio-cultural background of the L2 learners; sometimes they subsume part of the essay topics. Whenever in doubt, again, the concordance lines of the word combinations in question were examined. The results here indicate that these two proficiency-specific

subcorpora extracted from the Longman Learners' Corpus (LLC) somehow included many learner essays which were based on the same essay questions. Given that the source texts of LLC come from the teachers or students who voluntarily contributed their essays, it is not surprising to find that a great many of essays appear to come from perhaps dozens of writing classes.[16] Although during the initial stage of data selection, an attempt was made to avoid including too many essays which contained the same topics, the constraints of LLC data inevitably impose such possible skews towards context-dependent bundles.

The context-independent word combinations left, however, are not the finalised targets for investigation. There are some overlaps between bundles, which might contaminate the analysis results in some measure by inflating the counts of certain word combinations. In the next section, those overlapping word combinations will be categorised and discussed.

3.6.2 Data Deflation

During the process of examining bundles, it was found that sometimes two or three lexical bundles retrieved are actually part of a longer expression, and yet as the result of automatic retrieval, the longer expression is split into two or three shorter units for investigation. Overlapping word sequences (which were indicative of 5-word or even 6-word bundles) could inflate the results of quantitative analysis. Overlaps were manually checked via concordance analyses. For example, in BAWE-EN, *this may be due* and *may be due to* share a longer expression unit *this may be due to*, and examination of concordance lines indicate that these two four-word bundles which each occur four times originate from exactly the same four contexts. As a consequence, certain types of bundles can be undesirably over-represented when the quantitative analysis is conducted. In BAWE-EN, 36 lexical bundles out of the total

[16] It is not clear whether each learner contributed only one piece of written sample as this information is not recorded in LLC annotation.

120, over one quarter of them, are somewhat overlapping with another one or two lexical bundles. With regard to BAWE-CH and FLOB-J, there are 12 and 11 overlapping bundles respectively while the numbers of overlapping bundles in CEFR-C1 and CEFR-B2 writing are 12 and 27, each of which seems to be a substantial figure. A closer examination suggests that this issue is far more complex than it appears to be. Say one lexical bundle is part of a longer unit, thus overlapping with another lexical bundle. If all the occurrences in both bundles match perfectly such as in the case of *this may be due* and *may be due to* mentioned above, it would be easy to judge that the two split shorter bundles should combine into the five-word bundle *this may be due to* so as to avoid inflating results. However, if only some of the concordance lines of these two overlapping bundles are identical, then it is less easy to say whether the overlapping bundles should be combined and if so, whether the overlapping occurrences should be eliminated from quantitative analysis.

Nearly all of the overlapping bundles share the same structural and functional categorisations. If all the automatically retrieved word combinations are used for the quantitative analysis after the context-dependent bundles were filtered out, the result would still be contaminated because the overlapping bundles inflated certain structural and functional categories. In order to minimise the impact of inflation, a filtering system was devised to check against various conditions of overlaps. It was decided that frequency of occurrences should act as a decisive factor which determines if the overlapping bundles in question should be retained or not.

a) Complete Overlaps: 'Complete overlaps' refer to two four-word bundles which are actually derived from a single five-word combination. For example, *it has been suggested* and *has been suggested that* both occur six times, coming from the longer expression *it has been suggested that*. In this condition of complete one-to-one match, only the longer five-word units will be included in the finalised lexical bundles for investigation. There are four

such instances in Modular Study 1 and five in Modular Study 2 as illustrated below.

Table 3-11 Instances of 'Complete Overlaps'

Modular Study 1				
	Overlapping bundles	**Freq**	**Combined bundles**	**Freq**
FLOB-J	*can be seen that*	5	*it can be seen that*	5
	it can be seen	5		
	the turn of the	7	*the turn of the century*	7
	turn of the century	7		
BAWE-EN	*this may be due*	4	*this may be due to*	4
	may be due to	4		
BAWE-CH	*it has been suggested*	6	*it has been suggested that*	6
	has been suggested that	6		
Modular Study 2				
CEFR-C1	*it is not easy*	4	*it is not easy for*	4
	is not easy for	4		
CEFR-B2	*as a matter of*	4	*as a matter of fact*	4
	a matter of fact	4		
	the rest of the	4	*the rest of the world*	4
	rest of the world	4		
	there are a lot	11	*there are a lot of*	11
	are a lot of	11		
	will not be able	6	*will not be able to*	6
	not be able to	6		

b) Complete Subsumption: This type of overlap refers to a situation where two or more four-word bundles overlap and the occurrences of one of the bundles subsume those of the other overlapping bundle(s). For example, *as a result of* occurs 17 times while *a result of the* occurs five times which constitute a subset of the 5-word bundle *as a result of the*. Under such circumstances, it would be more sensible to keep only one of the overlapping bundles to avoid undesirably over-representing the same notions. Therefore, the overlapping word sequences of this kind, 13 cases in total, were combined into longer units so as to guard against inflated results (see Table 3-12). In the case of complete subsumption, a pair of brackets with the mark + was added in each finalised five-word combination to indicate the extended part of the longer unit.

Table 3-12 Instances of 'Complete Subsumption'

Modular Study 1				
	Overlapping bundles	**Freq**	**Combined bundles**	**Freq**
FLOB-J	*in the context of*	19	*in the context of+(the/a)*	19
	the context of the	4		
	the context of a	6		
BAWE-EN	*as a result of*	17	*as a result of+(the)*	17
	a result of the	5		
	this is due to	5	*this is due to+(the)*	5
	is due to the	4		
	one of the most	7	*one of the most+(important)*	7
	of the most important	4		
	to the fact that	8	*(due)+to the fact that*	8
	due to the fact	6		
	it can be seen	12	*it can be seen+(that)*	12
	can be seen that	11		
BAWE-CH	*an important role in*	5	*(played)+an important role in*	5
	played an important role	4		
	at the same time	24	*(and)+at the same time*	24
	and at the same	4		
Modular Study 2				
CEFR-C1	*it is obvious that*	11	*it is obvious that+(the)*	11
	is obvious that the	4		
	in such a way	5	*in such a way+(that)*	5
	such a way that	4		
CEFR-B2	*my point of view*	5	*(from)+my point of view*	5
	from my point of	4		
	the best way to	5	*(is)+the best way to*	5
	is the best way	4		
	the most important thing	7	*the most important thing+(is)*	7
	most important thing is	6		

c) **Partial Subsumption:** If the frequency of the longer combined unit, i.e. the five-word combination, is deducted from that of any of the overlapping bundles, and the frequency left is lower than four, this means that without the longer unit, the overlapping bundle in question would not be sustained since the cut-off frequency to determine a lexical bundle is four times in the current study. In such cases, only the more frequent bundle will be retained in the finalised set of lexical bundles for investigation so that the inflated results can be eschewed. The retained bundle would be represented with the overlapping part added in brackets as

Condition b), and the frequency of the retained bundle is the summed occurrences of the overlapping bundles minus the frequency of the extended five-word unit.

Some doubts might arise concerning whether this practice of deflation would cause under-representation of certain bundles because only one of the overlapping bundles entails some occurrences shared with the other bundle but the bundle with lower frequency would be eliminated. However, we have to carefully examine this issue from the perspective of structural and functional categorisation, which is one of the primary modes of analysis to be conducted in this thesis and will be discussed in the next chapter. In terms of bundle types, many of the overlapping components in the above instances are articles such as *the* or *a*, and the addition of these components to the retained bundles does not impose any undesirable consequences on either structural or functional categorisation. With respect to frequency, the fairest way to present the occurrence counts appears to be the sum of the two or three overlapping bundles minus that of the extended five-word units. That is to say, counting the frequency of this kind of overlaps is only fair when the overlapping part of frequency being repeatedly included is deducted, if we would like to better reflect the occurrences of such overlaps, and this is exactly the solution adopted here.

The overlaps of this type are listed in Table 3-13. The bundle below the dotted line in each row (i.e. each case of overlapping bundles) indicates the overlapped five-word unit. Again, a pair of brackets with the mark + was added in each finalised five-word combination to indicate the extended part of the longer unit.

Table 3-13 Instances of 'Partial Subsumption'

Modular Study 1				
	Overlapping bundles	**Freq**	**Combined bundles**	**Freq**
FLOB-J	*the end of the*	10	*(at)+the end of the*	*13*
	at the end of	6		
	at the end of the	3		
	the way in which	11	*the way in which+(the)*	*14*
	way in which the	5		
	the way in which the	2		
BAWE-EN	*is an example of*	5	*is an example of+(a)*	*6*
	an example of a	4		
	is an example of a	3		
	an example of this	7	*an example of this+(is)*	*8*
	example of this is	6		
	an example of this is	5		
	can be applied to	6	*can be applied to+(the)*	*7*
	be applied to the	5		
	can be applied to the	4		
	the end of the	10	*(at)+the end of the*	*13*
	at the end of	9		
	at the end of the	6		
	the development of the	11	*(for)+the development of the*	*13*
	for the development of	5		
	for the development of the	3		
	also be used to	4	*(can)+also be used to*	*5*
	can also be used	4		
	can also be used	3		
	can be used to	17	*(and)+can be used to*	*19*
	and can be used to	4		
	and can be used to	2		
BAWE-CH	*in the end of*	6	*in the end of+(this)*	*7*
	the end of this	4		
	in the end of this	3		
	at the end of	5	*at the end of+(the)*	*8*
	the end of the	4		
	at the end of the	1		
	is one of the	13	*is one of the+(most)*	*21*
	one of the most	13		
	is one of the most	5		

Modular Study 2				
	Overlapping bundles	**Freq**	**Combined bundles**	**Freq**
CEFR-C1	*as a matter of*	5	*as a matter of+(fact)*	6
	a matter of fact	4		
	as a matter of fact	3		
	at the beginning of	5	*at the beginning of+(the)*	6
	the beginning of the	5		
	at the beginning of the	4		
CEFR-B2	*have the right to*	14	*(can)+have the right to*	15
	can have the right	5		
	can have the right to	4		
	it is very difficult	6	*it is very difficult+(to)*	7
	is very difficult to	5		
	it is very difficult to	4		
	some people think that	4	*some people think that+(the)*	6
	people think that the	4		
	some people think that the	2		

It also has to be noted that after the combination of overlapping bundles, the structural categorisation for *(at)+the end of the* and *(for)+the development of the* would accordingly alter from the NP-based category to the PP-based category, which will be addressed in Chapter 4. The concordance lines, however, have been checked and confirmed that the collocation prior to *the end of* and *the development of* are indeed mostly prepositions in addition to *at* and *for* present here.

There are also a number of peculiar conditions of overlapping bundles which do not fit into any of the three major conditions described above. As the justification to combine or not combine these peculiar overlapping bundles sometimes appear to be rather trivial and lengthy, for the sake of space and clarity, it is decided to move the discussion of these peculiar instances to Appendix 1.

From the above accounts, it can be seen that the system outlined here is both methodologically and perceptually complex. It also has to acknowledge that this system does not fully resolve the problem of over-representation or under-representation in determining

lexical bundles. Other lexical bundles which have not been discussed are also very likely to be part of a longer expression. The reason why they do not emerge as overlapping lexical bundles is only because the frequency does not reach the cut-off frequency of four times. For example, the preposition *to* precedes three occurrences of the bundle *the rest of the*. With one more occurrence, a new four-word bundle *to the rest of* would come into being. Unfortunately, the approach of investigating lexical bundles fundamentally hinges on the set frequency, and it is impossible for the researcher to cater for all those bundle candidates that fail to reach the frequency threshold.

Generally speaking, among the overlapping lexical bundles, only the one with the highest frequency of occurrences would be used as the basis when types and tokens of lexical bundles are calculated for analysis. This practice is considered to be justifiable in the sense that by removing the overlapping bundles with lower frequency counts, the risk of inflated results can be effectively decreased since the overlapping bundles virtually express the same notion. On the other hand, these removed bundles would be represented as the extended parts in brackets in the retained bundles, which could still be traced if needed.

The numbers of lexical bundles before and after this deflation procedure are presented in Table 3-14. As can be observed with regard to the fact that the number of bundles in native British students' writing reduced drastically after the process of deflation, a possible explanation is that British students tend to use expressions of longer length and thus generate more overlapping four-word bundles than the other two groups of writers. At the very early stage of bundle retrieval, the fact that native peer writing entails the most lexical bundles among the three groups of writers appears to be fairly intriguing. However, after the filtering process, there seems to be a sort of pattern with writing competency of the three groups of writers and numbers of lexical bundles in Modular Study 1. It is likely that the less competent writers are, the fewer lexical bundles will appear in their writing. The assumption above

could lead to the implication that the competent writers might have more formulaic expressions at their disposal. This assumption, however, is challenged by the result revealed in Modular Study 2. The interacting relationship between writing proficiency and this frequency approach will be discussed in Chapter 7.

Table 3-14 Number of lexical bundles before and after filtering and deflation in the corpora investigated (types)

Study	Corpus	Raw instances	Filtered instances (Context-dependent bundles removed)	Deflated instances (finalised bundles)
Modular Study 1	BAWE-CH	90	86	80
	BAWE-EN	120	120	104
	FLOB-J	118	116	108
Modular Study 2	CEFR-B2	164	84	71
	CEFR-C1	169	42	37

Through such a manual examination of removing context-dependent bundles and combining overlapping bundles, the repertoire of bundles in each subcorpus was substantially downsized (see Table 3-15), particularly for the subcorpora from Modular Study 2. Although this was not expected when the cut-off threshold was considered to retrieve a good number of bundles for investigation, at least now we have more confidence in the quality of those 'cleaned' bundle data as the impact from those undesired variables which come along with this frequency approach have been decreased.

Table 3-15 Number of bundles before and after the removal of context dependent bundles and overlaps

Study	Corpus	Before refinement		After refinement	
		No. of lexical bundles (types)	No. of lexical bundles (tokens)	No. of lexical bundles (types)	No. of lexical bundles (tokens)
Modular Study 1	BAWE-CH	90	554	80	507
	BAWE-EN	120	757	104	667
	FLOB-J	118	749	108	704
Modular Study 2	CEFR-B2	164	936	71	411
	CEFR-C1	169	996	37	241

3.7 Conclusion

This chapter started with a review of the literature on phraseology and then focused on the studies which adopted this frequency-based approach to defining phraseological units. Such an overview underpins the theoretical and operational frameworks for this project, e.g. the determination of frequency and dispersion thresholds. After repeated experiments with the corpus data used in this thesis, however, it was found that the conventional way of reporting a normalised cut-off frequency could be misleading. Here it is argued that both the normalised and the raw frequency thresholds should be reported to genuinely reflect the relationship between the converted frequency threshold and corpus size. Meanwhile, during the process of examining retrieved clusters, two conditions were found which could undermine the validity of analysis results: context-dependent bundles, and overlapping ones. For the former, those bundles which form part of the essay questions or relate to the socio-cultural contexts where the bundles occur, usually with proper nouns, were manually removed. For the latter, a system was devised to categorise various types of overlaps and deal with them accordingly. The final 'cleaned' bundles are considered to be of higher quality for follow-up analysis as they represent the true building blocks of discourse which this thesis aims for.

In the next chapter, the second part of the methodology in this thesis, i.e. ways of comparison and classification of bundles, will be described.

Chapter 4 Analytical Framework

This chapter deals with the second framework underpinning the analysis, i.e. the analytical framework. It will begin with the rationale behind the types of comparisons made in Chapters 5 and 6, which involve type and token comparisons, keyness analysis, and analyses of structural and functional distributions in two modular studies. What follows is the illustration of structural and functional categorisation, which will start with a review of Biber et al's taxonomy and then move on to the issues involved in applying the existing taxonomy for bundle categorisation, and how the taxonomy could be modified to better accommodate the data used in this project. In other words, in addition to summarising how different types of comparisons would be carried out in this thesis, this chapter also deals with the second methodological/procedural research question (see Section 3.1) which aims to improve the current taxonomy established by Biber et al. (1999, 2003, 2004, 2007) for classifying bundle structures and functions in order to create a consistent and robust categorisation scheme.

4.1 Ways of Comparing Corpora

4.1.1 Two Modular Studies

An ideal scenario of comparison which includes L1 expert writing, L1 novice writing, and L2 writing of different proficiency levels is to have all the native and non-native speakers respond to the same writing task under an identical setting so that proficiency would be the only variable that affects the quality of writing. In such an ideal scenario, researchers would have more confidence in searching for developmental patterns as they could assume that L1 expert writing is placed at one end of the linear relationship of proficiency and the weakest L2 writing at the other end. In reality, it is difficult to have a large number of subjects (such as nearly 600 written samples investigated in this thesis including native and non-native

writing at various levels) finish a writing task under the same experimental setting. And even if such data was collected, questions about authenticity would be raised as it would not be considered 'naturalistic' (yet note that the Longman Learners' Corpus contains non-naturalistic as well as naturalistic data). Another issue is that we cannot ascertain that every subject invited to this experimental setting would exactly produce a piece of writing corresponding to his/her supposed language proficiency (see Sections 2.2 and 2.3 for the discussion of proficiency). In the light of accessibility of corpus data and comparability, it was then decided to separate the available corpus data into two modular studies. The first study deals with writing produced in an academic context, and the second study concerns rated learner essays termed as EAP-like writing. The key notions in both studies are summarised below.

The first modular study falls within the scope of conventional learner corpus research, which compares native language and learner language. Yet it also distinguishes itself from most of the literature in the sense that both native expert writing and native novice writing are included and the genres it adopts is rather different from general learner corpus studies. As discussed in Section 2.1.2, a great many learner corpus studies compare learner writing with native novice writing (usually produced by British and American students). Take ICLE and LOCNESS data for example. Only argumentative essays are included in these corpora, which were presumably produced as a writing exercise or take-home assignment under the setting of a language course.[17] The purpose of writing such an essay is probably to polish the student's writing skills. The major focus of such essay writing is the language itself (e.g. grammar, lexis, coherence and cohesion, etc.), and the nature of such data is more experimental rather than naturalistic. In comparison, the written samples used in the first study in this thesis are

[17] See the guidelines for collecting subcorpus data on the ICLE website (http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm, visited on 10 June, 2009).

all naturalistic data generated under real-life academic contexts either as part of profession requirement (for academics) or part of degree requirement (for university students). The native expert writing, the component of academic prose extracted from the FLOB corpus (FLOB-J) consists of extracts of journal papers or academic books. The two groups of student writing, L2 writing of L1 Chinese students (BAWE-CH) and L1 peer writing of British students (BAWE-EN) both come from the BAWE corpus, which collected proficient assessed texts from British universities including essays, critiques, proposals, and among others (Alsop & Nesi, 2009). The major focus of such academic writing is the content of texts (originality, methodology, criticalness, contribution, etc.) as opposed to the language per se, although language use might play some role in its quality. More detail about data selection for the investigation of academic writing will be addressed in Chapter 5.

The second modular study comes under second language development research which compares learner language across proficiency levels. The argumentative and expository writing as well as some academic essays produced by L2 learners of L1 Chinese are extracted from the Longman Learners' Corpus. As the text types are similar to those targeted at students who are learning English for Academic Purposes (EAP), which aims to prepare L2 learners for academic studies at the tertiary level, they are termed as 'EAP-like writing' in this thesis. There are two reasons why these types of essays are included. The first is that academic writing generally involves exposition as well as argumentation so that this allows comparability to some extent. The other reason, a practical one, is that there would not be enough texts if only argumentative essays are used as the general practice in most learner corpus studies. After being double-marked with the rating scale from the Common European Framework of Reference (CEFR), there was not enough data from the top and the bottom groups, a challenge that many L2 developmental studies are confronted with. Only two levels of writing contained a sufficient number of texts for analysis: CEFR-B2 and CEFR-C1. The

details about rating and data selection will be described in Chapter 6.

At the beginning of my research, it was planned to use authentic examinees' scripts from a large-scale English proficiency test in Taiwan. The data that I was promised, however, failed to appear. For a pilot analysis, I used learner data from the Longman Learners' Corpus and compared it with the academic prose from FLOB and some linguistic essays written by British students that I had collected myself, and thereby the issue of genre comparability was raised in my upgrade meeting. Given the accessibility of data, it was therefore decided to split the comparisons into two modular studies in order to remedy this problem. Yet an overall discussion of results from both modular studies (Chapters 5 and 6) will be presented in Chapter 7 in the sense that quite a few patterns of learner idiosyncrasies are found across groups of learner writing regardless of the proficiency and genre differences.

4.1.2 Types vs. Tokens

A potential problem with comparing lexical bundles across corpora involves what is actually counted. This is particularly relevant for the distinction between types and tokens. Should we count the number of *types* of bundles (e.g. counting *as a result of* and *it is possible to* as each one type of bundle), or should we just count the total occurrences of bundles (e.g. *as a result of* might occur 20 times in one corpus and 50 times in another corpus)? One corpus could exhibit a very narrow range of types of bundles but have very high frequencies of them, while another could have the opposite pattern. To take the issue of range and frequency into account, I have carried out two sets of analyses, first looking at different types of bundles (*types*) and then examining overall frequencies of bundles (*tokens*).

As discussed in Chapter, 3, this type/token distinction is also important in that the number of bundle tokens increases with corpus size whereas the number of bundle types does not have such a simple linear relationship with corpus size. In theory, the number of bundle types we have stored in our mental lexicon is finite, just as our vocabulary is finite. Yet it is

still unclear what the optimal corpus size should be in order to retrieve the maximum number of bundle types. These issues will affect the quantitative comparison of corpora with different sizes, which will be discussed in Chapter 7.

4.1.3 Keyness Analysis

As mentioned in Section 1.2, *WordSmith* (Scott, 2007) can identify the words/clusters which are significantly more or less frequent in a target corpus when compared with a reference corpus by means of statistical tests (chi-square or log-likelihood) which compare the frequencies of a lexical unit (words or clusters) in both corpora, taking into account the overall size of each. A 'keyness' score is given for each of the words/clusters that has statistically significant difference in frequency between the two corpora. The higher the keyness score, the more statistically significant the key lexical unit. Here the current study only deals with selected clusters, i.e. lexical bundles, and I have therefore re-termed this approach as a 'keyness analysis' to avoid confusion.

A keyness analysis takes into account both corpus size and frequency, which is considered to be important for the present study in the sense that a defined lexical bundle, a four-word sequence which reaches a set frequency and dispersion threshold, actually fails to reflect any statistically significant difference of the frequency counts between corpora compared. A keyness analysis, however, can provide a remedy for this problem by virtue of signposting the word combinations which occur significantly more or less frequently in comparing two corpora. The results of keyness analysis will be presented and discussed after quantitative and qualitative analysis in Chapters 5 and 6. A keyness analysis which uses native expert writing as the reference corpus to be compared with the other four groups of non-expert writing will also be presented in Chapter 7, in which any 'key' bundles being overlooked earlier will be discussed too.

4.1.4 Structural and Functional Distribution

It has been found that lexical bundles do not carry a distinct linguistic function but rather are 'basic building blocks of discourse' (Biber et al., 2004, p. 371), which cover a wide range of diverse linguistic categories. The advantages of adopting such a corpus-driven methodology are two-fold (for the distinction between 'corpus-driven' and 'corpus-based' approaches, please see Section 1.1). First, without a pre-existing inventory of linguistic devices as the target(s) for investigation, this methodology relies on computerisation to produce a list of multi-word sequences that occur over a specified frequency and distribution. This bottom-up approach hence allows a more thorough examination of learner language, and any problematic linguistic aspects that might be implicit otherwise can be revealed. The second advantage stems from the dramatic difference between written and spoken registers in terms of structures and functions of lexical bundles uncovered by L1 studies (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber et al., 2003; Biber et al., 2004). This previous finding concerning distinctive characteristics between registers can be used to consolidate the general assumption throughout the present study: the more proficient learners are, the more native-like their writing is inclined to be, and by contrast the writing in the less proficient learners is prone to be more conversation-like, as has been discussed in Chapter 1 detailing the Research Questions. This thesis hence also takes advantage of the structural and functional taxonomy developed by Biber and his colleagues (hence it is also a top-down 'corpus-based' approach), with some minor modifications when applying their taxonomy with the FLOB, BAWE, and LLC data described earlier. This process will be addressed in detail in the following two sections.

4.2 Structural Classification

4.2.1 Background

Biber et al's work on lexical bundles in the Longman Grammar of Spoken and Written English (1999, pp. 997-1025) is one of the pioneering studies in this area, and its framework of structural classification has often been used in other studies on lexical bundles since then (Cortes, 2002, 2004; Hyland, 2008a, 2008b). As pointed out by Biber et al. (*ibid*), although very few lexical bundles form a complete structural unit by themselves, most bundles have strong grammatical correlates, and thus it is possible to make categorisations on this basis. In the Longman Spoken and Written English (LSWE) corpus, lexical bundles were grouped into fourteen major categories in conversation and twelve major categories in academic prose with some overlaps between the categories. In order to examine the feasibility of the Longman framework with the relatively small-scale data in the present study, a preliminary pilot structural classification was carried out with the lexical bundles retrieved from the corpora used in the first modular study: BAWE-CH, BAWE-EN, and FLOB-J. The result was then compared with the proportions of grammatical categories in the LSWE Corpus. Despite the drastic difference in the corpus size[18] and the different frequency thresholds (ten times per million words for LSWE and around 25 times per million words for FLOB-J, BAWE-EN and BAWE-CH), the result shows a surprising match between the academic prose component of LSWE and FLOB-J while the proportions in the two groups of student writing fluctuate to some extent when compared with the academic prose in LSWE (see Table 4-1). Not only does such preliminary comparison lend a good deal of credence to the use of smaller corpora with different frequency cut-offs in the current project, but it also indicates a gap between

[18] In LSWE, the spoken data reaches almost 4 million words, and the register of academic prose is as large as 5.3 million words. The average corpus size for Modular Study 1 is approximately 150,000 words.

native expert academic prose and immature student academic writing. This gap might be a result of genre difference between published academic essays and university assignments, but it is also possible that it hinges upon writing proficiency. Therefore, this LSWE framework of structural categorisation is employed as a starting point. Whether this framework requires some modification will be further explored with the corpus data in the present study.

From the patterns emerging from Table 4-1, it is also found that two grammatical categories more prevalent in LSWE conversation, (2) 'copula *BE* + NP/AdjectiveP' and (3) 'VP with active verb', show a higher proportion of use in BAWE-EN and BAWE-CH but not in FLOB-J. This may suggest that immature student writing tends to be more similar to speech – a hypothesis that will be returned to later in the thesis.

Table 4-1 Proportional distribution of four-word lexical bundles across the major structural patterns in LSWE, FLOB-J, BAWE-EN, and BAWE-CH (adapted from the Longman Spoken and Written English (Biber et al., 1999, p. 996, with the addition of the three academic corpora used in this thesis)

		CONV (LSWE)	ACAD (LSWE)	FLOB-J	BAWE-EN	BAWE-CH	Example
	patterns more widely used in conversation						
(1)	personal pronoun + lexical verb phrase (+ complement clause)	44%	--	--	--	--	*I don't know what*
(2)	COPULA *BE* + NP/AdjectiveP	8%	2%	2.6%	10.6%	6.3%	*is one of the*
(3)	VP with active verb	13%	--	0.9%	2.9%	6.3%	*has a number of*
(4)	*yes-no* and *wh*-question fragment	12%	--	--	--	--	*can I have a*
(5)	(verb +) *wh*-clause fragment	4%	--	--	--	--	*know what I mean*
	patterns more widely used in academic prose						
(6)	noun phrase with post-modifier fragment	4%	30%	32.5%	15.4%	15%	*the nature of the*
(7)	preposition + noun phrase fragment	3%	33%	36%	28.8%	32.5%	*as a result of*
(8)	anticipatory *it* + VP/adjectiveP + (complement-clause)	--	9%	8.8.%	5.8%	8.8%	*it is possible to*
(9)	passive verb + PP fragment	--	6%	7%	10.6%	5%	*is based on the*
(10)	(VP +) *that*-clause fragment	1%	5%	2.6%	4.8%	6.3%	*should be noted that*
	patterns used in both registers						
(11)	(verb/adjective +) *to*-clause fragment	5%	9%	7%	18.3%	15%	*are likely to be*
(12)	other expressions	6%	6%	2.6%	2.9%	5%	*as well as the*
	Total	100%	100%	100%	100%	100%	

4.2.2 Issues Concerning Structural Categorisation

This section discusses the major structural categories which were applied in the corpus data for this project and the rationale behind it. A number of issues accompanying the application of classification such as ambiguity in categorisation and the possible partial overlaps in these computer-derived word sequences are also addressed.

Biber et al. (2003) and Biber et al. (2004) both mentioned a phrasal-clausal distinction in their classification results of bundle structures. In the 2003 paper, they stated that the majority of lexical bundles in academic prose are noun phrases with post-modifiers (e.g. *the nature of the*) and prepositional phrase fragments (e.g. *as a result of the*), thus being mostly 'phrasal rather than clausal' (p. 77). In their 2004 paper, they again referred to the bundles which incorporated noun and phrase fragments as being 'phrasal' and bundles which contained verb phrase fragments (e.g. *have a look at*) or dependent clause fragments (e.g. *that there is a*) as having 'clausal components' (p.380). Their description appears to imply that phrasal bundles only contain noun and prepositional components. Instead, as long as there is a verb component, no matter if it is a VP fragment or incorporated in a dependent clausal fragment, then the bundle in question has a clausal component. This simplified dual distinction, however, overlooks the distinction between 'phrases' and 'clauses': a phrase is a cluster of words without either a subject or a verb or without both, whereas a clause has a subject and a verb. Given that phrases are not exclusive for word sequences in which a noun or a preposition functions as head, such use of umbrella terms in Biber et al's studies—phrasal/clausal—can be misleading in a sense as there are verb phrases (e.g. *play an important role*) as well as noun phrases and prepositional phrases. In terms of clauses, according to Quirk et al. (1972, pp. 64-65), most non-finite clauses do not have an overt subject (e.g. *It was Kim's idea to invite them all.*), and there are even verbless clauses (e.g. *He was standing with his back on the wall.*). The distinction between clauses and phrases then

appears to be blurred to a large extent. When putting the structural classification into practice for the corpora under investigation here, it was felt in the beginning that an upper layer, a phrasal-clausal distinction, should be added to the original LSWE framework so as to facilitate a comparison from a broader perspective. Nevertheless, the phrasal-clausal distinction used by Biber et al. does not seem to be able to stay intact through such examination. Since it is not the intention for this thesis to explore the difference of definition between a phrase and a clause, this dual distinction is thus discarded.

Biber et al. (2004) also proposed a three-way distinction for structural classification, i.e. VP-based bundles (e.g. *is going to be*), dependent-clause-based bundles (e.g. *I don't know if*), and NP/PP-based bundles (e.g. *one of the things* and *at the end of*), although they described the first two structural types as being clausal and NP/PP-based bundles as being phrasal. The dividing line between phrasal and clausal word sequences has been discussed and would no longer be applied in the current study. With regard to VP-based lexical bundles and bundles that incorporate dependent clause fragments, since they both have verb components in the definition, and dependent-clause-based bundles occur rarely within academic writing, there appears no need to maintain a distinction between these two. Thus far, based on the analysis of my data, three major categories have been finalised: NP-based bundles, PP-based bundles, and VP-based bundles, and there are subcategories under each major category, which will be further discussed.

Another complex issue arising when applying the classification system is that some of the bundles are eligible to be assigned into more than one corresponding category on the basis of its surface structure. As Biber et al. noted, 'these [structural] categories are not always mutually exclusive' (1999, p. 1001). Such ambiguous overlapping occurs most often in '*that*-clause fragments' or '*to*-clause fragments' (Categories (10) and (11) in Table 4-1) with other verb-based subcategories such as 'passive verb + prepositional phrase fragment'

(Category (9)), 'anticipatory *it* + verb phrase/adjective phrase' (Category (8)), or 'verb phrase with active verb' (Category (3)). For example, *it is clear that* can be allocated to both 'anticipatory *it* + verb phrase/adjective phrase' and '*that*-clause fragment' patterns. Additionally, *to ensure that the* can fall in the categories of '*to*-clause fragment' or '*that*-clause fragment' while *to be added to* can be assigned to '*to*-clause fragments' or 'passive verb + prepositional phrase fragment'. Within the LSWE structural categories created by Biber et al. (1999), an unwritten rule appears to govern the priority of classification. The impression is that the grammatical categorisation seems to have been determined by the first one or two words in each bundle. Therefore, *it is clear that* falls in the category of anticipatory *it* pattern while *to ensure that the* is classed as a '*to*-clause fragment' rather than a *that*-clause pattern, and *to be added to* is put in the category of '*to*-clause fragment' instead of 'passive verb + preposition phrase fragment'. As a matter of fact, the nature of lexical bundles bears on this classification issue. Lexical bundles are purely the products of computation regardless of the completeness of grammatical status. Most four-word bundles have been found to bridge two structural units. In the case of a composite lexical bundle, i.e. a bundle consisting of more than one structural unit, the first unit would dominate the second unit in terms of structures. Considering that in the English syntax system, an embedded structure such as a clause is generally introduced by the word or phrase prior to it, grouping the lexical bundles according to some aspect of word order within each lexical bundle, wherever ambiguity occurs, appears sensible. The practice of this argument can be exemplified through the classification of lexical bundles in Table 4-2 below:

Table 4-2 Some examples of composite lexical bundles and their structural units

Category (the first structural unit)	(The second structural unit)	**Examples of Composite Lexical Bundles**
prepositional phrase +*of*-phrase fragment	(noun phrase fragment)	*as part of the, in terms of a, in view of the*
other prepositional phrase	(*that*-clause fragment)	*by the fact that, in the sense that*
verb phrase with active verb	(*that*-clause fragment)	*bear in mind that*
anticipatory *it* + verb phrase /adjective phrase	(*that*-clause fragment) (*to*-clause fragment)	*it is believed that, it is clear that* *it is difficult to, it is important to, it is necessary to, it is possible to*
that-clause fragment	(existential *there* + *BE* pattern fragment)	*that there is a, that there is an, that there is no*
pronoun/noun phrase + BE/verb	(*that*-clause fragment)	*there is evidence that, we can see that*
to-clause fragment	(passive verb + prepositional phrase fragment) (*that*-clause fragment)	*to be added to* *to ensure that the*

* The underlined part indicates the first structural unit in a lexical bundle.

Although the above principle caters for most of the ambiguous cases occurring in classification, it is not completely indisputable. To begin with, it is not always possible to categorise a word combination simply on the basis of the first structural unit in it. There is only one category 'passive verb + prepositional phrase fragment' (Category (9) in Table 4-1) for word combinations incorporating the component of passive verbs, for example, as the majority of them are followed by prepositional phrase fragments and it does not appear too economical to create several categories simply to accommodate variations of bundles containing a passive verb. Embracing the LSWE framework, therefore, means that '*can be seen*' followed by different second structural units would be assigned to different categories although they might look fairly similar to each other at first sight, e.g. *can be seen in* allocated to the category of 'passive verb + prepositional phrase fragment', *can be seen that* being in the category of 'verb phrase + *that*-clause fragment', and *can be used to* being in the category of 'verb phrase + *to*-clause fragment'.

Another drawback of the structural classification directly involves the automatic approach of retrieving recurrent word combinations, thereby triggering the procedure of data deflation described in Chapter 3. Longer word sequences may break into two or more short word combinations. Distinctive categories could thus be assigned to two or more associated lexical bundles when they are in fact part of a longer unit of expression. For example, *a result of the* extracted from one student writing corpus is accordingly allocated into the category of noun phrase fragments while *as a result of*, also retrieved from the same corpus, falls in the category of prepositional phrase fragments, although all the five instances of the former actually incorporate with the latter forming the longer unit *as a result of the*.[19] Other intriguing instances were found when the repertoire of retrieved word sequences was carefully cross-checked, with particular attention to those that might incorporate another bundle into a longer unit. The results show that there exist a lot more bundles than expected that are actually part of longer word sequences. Take *the end of the* and *at the end of* for example. They both appear in the corpora investigated and in the LSWE corpus and thus are highly frequent expressions in academic written texts. Having checked the concordance lines and taken notes of the collocated words or phrases following *at the end of* and prior to *the end of the*, it was found that because the five-word word sequence *at the end of the* is shared by the two four-word bundles, the frequency counts have been inflated to some extent. That is to say, the same longer recurrent word sequences have been broken down into two lexical bundles as a result of the retrieval technique and hence have been included repetitively, either partly or completely. Focusing on the data extracted from one student writing corpus BAWE-EN, the five-word sequence *at the end of the* accounts for six occurrences in *at the end of* and *the end of the*. Without the six instances, there are only three occurrences left in *at the end of*

---

[19] The example illustrated here, a pair of overlapping bundles (*a result of the* and *as a result of*), has been combined into one longer unit *as a result of+(the)* (see Section 3.6.2).

collocated with other words, and they alone do not meet the cut-off frequency, four times set by the current study (see Table 4.3 for this overlapping phenomenon of *at the end of* and *the end of the* in the first modular study).

Table 4-3 Collocation and frequency in *at the end of* and *the end of the*

Bundle	Corpus	Freq	Text	Collocated word/phrase & frequency
at the end of+*	BAWE-CH	5	5	the 1, observation 1, year 1, each 1, term 1
at the end of+	BAWE-EN	9	6	the 6, year 2, line 1
at the end of+	FLOB-J	6	5	the 3, this 1, these 1, travel 1
+the end of the	BAWE-CH	4	4	at 1, in 1, by 1, of 1
+the end of the	BAWE-EN	10	8	at 6, by 3, toward 1
+the end of the	FLOB-J	10	9	at 3, towards 2, by 1, before 1, as from 1, it was 1, Ø#1

* The '+' mark indicates that the lexical bundle in question overlaps with another bundle, and its position shows where the overlap is.

This symbol Ø indicates that 'the end of the' is the start of a sentence.

The frequency inflation may not appear critical enough here, but some extreme cases were also found. For example, there were five occurrences respectively of *it can be seen* and *can be seen that* in FLOB-J, and both sets of bundles are retrieved from exactly the same instances of a longer expression *it can be seen that*. Other examples with high inflating frequency counts are *be argued that the/could be argued that*, *it has been suggested/has been suggested that*, *in the context of/the context of the/the context of a*. Such repetitive inclusion was mostly discovered in the categories 'noun phrase + post-modifier fragment', 'prepositional phrase fragment', 'anticipatory *it* +VP/adjectiveP', and '*that*-clause fragment'. As a consequence, the counts of bundles in these categories would undoubtedly be inflated to some extent if an analysis is carried out comparing the structural difference across corpora. Such overlapping has been observed in Biber et al. (1999), which is marked with '+' either before or after the bundle in question to indicate the extension. For example, *it should be*

noted+ indicates an extended five-word bundle *it should be noted that+*, which in turn indicates a even longer six-word extension *it should be noted that the*. Bundle overlapping appears to be a prevalent phenomenon in the LSWE data of academic prose as 23 out of 69 bundles in the category of 'Noun phrase with *of-* fragment' alone are marked with one or two overlaps. Yet surprisingly, the researchers did not seem to take any precautionary measures to tackle this issue, which is probably because it would take a tremendous amount of time and effort to examine tens of thousands of concordance lines to see the extent of inflation in the LSWE corpus. As far as the present project is concerned, the small-scale data investigated does not hinder such close scrutiny. Furthermore, considering the complexity of the relationship between frequency counts, lexical bundles in question, and the longer unit shared, a deflation scheme was devised with an attempt to eliminate the phenomenon of overlap and obtain a fairer result. Although this scheme has been discussed in detail previously in Section 3.6.2 which addresses data deflation, it was not until the structural categorisation was being conducted with concordance checks that the problem of overlapping came to light. Categorisation and data deflation, therefore, were carried out virtually at the same time. The deflation procedure, however, does not completely solve this problem of overlaps. Indeed, although the major overlaps (e.g. *it can be seen* and *can be seen that* mentioned earlier) have been combined, it is not easy to deal with 'minor' overlaps. For example, in BAWE-CH, there is only one concordance line shared between one lexical bundle *at the end of+* with a frequency of five times and another *+the end of the* with a frequency of four times. Bundles with such slight overlaps probably only have some marginal effect on the analysis, and thus they are both retained in the finalised BAWE-CH bundle repertoire. Yet the issue of overlaps reveals that the structural categorisation and the deflation scheme can only deal with the superficial grammatical structures demonstrated in the four-word combinations. The syntactical status at the clause/sentence level and the context where the bundles occur (i.e. the

collocates prior to and following the bundles), unfortunately, cannot be taken into account. Therefore, *on the other hand* or *at the same time* would be categorised as PP-based bundles although they function as adverbials in the syntactical level. The bundle *the end of the* in BAWE-CH, therefore, would be grouped under the category of NP-based bundles although it is usually collocated with a preposition prior to it (see Table 4-3).

The abovementioned caveats might somewhat undermine the validity of results drawn from a classification system; nevertheless, they are by no means the by products of any classification frameworks. Instead, these issues touch on the nature of lexical bundles—they are recurrent word sequences retrieved from automatic computation regardless of structural completeness, which has an impact to some extent on how to structurally classify the word combinations. For example, noun phrase fragments are very often preceded by a preposition as exemplified by the case of *the end of the* and *at the end of* (*cf.* Table 4-3), and all the prepositional phrase fragments always have a noun following the initial preposition. It was thus once considered whether the separation between the categories of noun phrase fragments and prepositional phrase fragments is necessary or not in the light of the seemingly blurring boundary of these two categories. However, considering the comparability with other studies and the overall design of categorisation system, it was still decided to keep the three major categories: NP-based, PP-based, and VP-based bundles, and the analyses conducted with the three-way distinction proved to be fairly effective, which will be presented in Chapters 5 and 6.

Given all the issues discussed above, one has to bear in mind that the data have to be treated with caution and that the interpretation of quantitative analysis has to be made acknowledging these caveats.

4.2.3 A Modified System for Structural Classification

To sum up, all the above discussions formed the supplementary notes below for the structural classification system proposed by Biber et al. (1999). These points are underlying in their framework but not articulated.

1. If a lexical bundle is a complete structural unit such as *the right hand side*, then the categorisation is determined by its overall structural status.
2. If a lexical bundle bridges two structural units, the structural classification would be based on the first structural component if a corresponding category is available. For example, *it is possible to* would be allocated to the category of 'anticipatory *it* + verb/adjective phrase' instead of '*to*-clause fragment'.
3. Overlap of lexical bundles is a prevalent phenomenon which has to be dealt with to avoid bundle inflation, particularly in certain structural categories (see the deflation procedure in Section 3.6.2). Although the overlaps have been noted in Biber et al.'s (1999), no solution was proposed to prevent the undesired inflation of data.

In order to solve the problem of spanning across double grammatical categories, the option of categorising lexical bundles structurally with a hierarchical system was considered in the beginning. However, this idea was abandoned later because of two reasons. First of all, the framework established by Biber et al. (1999) is fairly thorough with fourteen categories for conversation and twelve categories for academic prose. As long as the classification procedure can be carried out consistently for the current study, and there is good justification wherever ambiguity occurs as having discussed earlier, it appears sensible to embrace an existing classification system which has been well applied rather than devising a new one.

Bearing in mind the above issues, the present study grouped the target word combinations into 13 subcategories which largely correspond to the structural categories for

bundles in LSWE academic prose but with some modifications to accommodate certain bundles due to learner idiosyncrasies. These subcategories are created under three major groups: NP-based, PP-based, and VP-based bundles. Note that the bundles presented here have been filtered through the procedures of context independence checks and the deflation system described in Chapter 3, thereby being a comprehensive finalised repertoire of lexical bundles investigated. If one lexical bundle appears in more than one corpus and has an extended part marked with brackets, such as *one of the most+(important)*, in one of the corpora, then both the original four-word form and the extended form are recorded here. If one bundle overlaps with another but they are not combined because the amount of actual overlap is too small (i.e. shared concordance lines lower than cut-off frequency four times), then a '+' marked is added either preceding or following the bundle in question to indicate the position of overlap. For the marking scheme of overlapping bundles and individual overlapping cases, please refer to Section 3.6.2 and Appendix 1.

NP-based Bundles

A. Noun phrase with *of*-fragment (NP+*of*)

- a function of the, a function of time, a great deal of, a great number of, a high level of, a large amount of, a number of factors, a wide range of, an example of this, an integral part of, and the use of, end of the spectrum, most of the people, one of the main, one of the most, one of the most+(important), per cent of the, the creation of a, the development of the, the end of the, the existence of a, the history of the, the impact of the, the importance of the, the length of the, the magnitude of the, the nature of the, the point of view, the quality of the, the rest of the, the rest of the world, +the result of the, +the result of this, the results of the, the role of the, the rules of the, the second half of, the size of the, the status of the, the strength of the, the structure of the, the top of the, the turn of the century, +the use of the, the value of the

B. Noun phrase with other post-modifier fragment (NPf)

- a lot of people, +a lot of problem(s), a lot of time, a very important role, and a lot of, people who live in, the degree to which, the extent to which, the fact that the, the fact that they, the fact that this, the relationship between the, the right hand side, the way in which, the way in which+(the), the ways in which

PP-based Bundles

C. Prepositional phrase with embedded *of*-phrase fragment (PP+*of*)

- (at)+the end of the, (for)+the development of the, (from)+my point of view, as a function of, as a part of, as a result of, as a result of+(the), as a way of, as one of the, as part of a, as part of the, as the result of+, at each end of, at the beginning of, at the beginning of+(the), at the end of+(the), at the expense of, at the heart of, at the time of, by a variety of, by the presence of, for each of the, for the development of, for the use of+, in a number of, in terms of a, in terms of its, in terms of the, in the absence of, in the case of, in the context of, in the context of+(the/a), in the course of, in the end of+(this), in the face of, in the form of, in the hands of, in the light of, in the number of, in the presence of, in the process of, in view of the, of a number of, of some of the, of the number of, of the number of, on a number of, on the basis of, on the part of, to that of the, to the development of, with the addition of, with the development of, with the introduction of

D. Other prepositional phrase fragment (PPf)

- (and)+at the same time, (due)+to the fact that, all over the world, and as a result, as a matter of+(fact), as a matter of fact, at the same time, by the fact that, for a long time, in addition to the, in an attempt to, in contrast to the, in more detail in, in relation to the, in so far as, in such a way+(that), in the first place, in the following paragraphs, in the long run, in the recent years, in the same way, in the sense that, in this essay I, on the one hand, on the other hand, through the use of, to a certain extent, to a lack of, to a large extent, to the fact that, with respect to the

VP-based Bundles

E. *it* + verb phrase/adjective phrase (*it* pattern)

- it can be argued+(that), it can be seen, it can be seen+(that), it can be seen that, it could be argued that+(the), it has been suggested, it has been suggested that, it has not been, it has to be, it is a good, it is a very, it is also a, it is believed that, it is clear that, it is difficult to, it is easy for, it is easy to, it is estimated that, it is hard to, it is important to, it is necessary to, it is not a, it is not always, it is not clear, it is not easy+(for), it is obvious that+(the), it is possible to, it is true that, it is true that, it is very difficult, it is very difficult+(to), it should be noted, it would have been

The original category in the LSWE taxonomy is 'anticipatory *it* + verb phrase/adjective phrase'. Yet many of the structures starting with *it* found in learner writing, particularly CEFR-B2 writing, such as *it is a very* or *it is also a*, are pronouns followed by a predicate rather than the 'anticipatory *it*' pattern. This category is thus modified, with the word 'anticipatory' being removed. This difference of usage of this *it* pattern across groups of writers will be discussed in Chapters 6 and 7.

F. (Verb phrase+) *that*-clause fragment (*that* clause)

- Verb phrase + *that*-clause: bear in mind that
- *That*-clause: that there is a, that there is a/an, that there is no

The original LSWE category is '(Verb phrase) + *that*-clause fragment'. Interestingly, the first subgroup with a verb phrase plus a *that*-clause fragment are intended for the bundles with some original four-word forms including '*+be argued that the*', '*,+can be argued that*', '*+can be seen that*', '*+could be argued that*+', '*+has been suggested that*'. Yet these instances have all been combined with the anticipatory *it* pattern and thus categorised under '*it* + verb phrase/adjective phrase'. The only exception which does not fit into this pattern is *bear in mind that*.

G. (Verb/adjective +) *to*-clause fragment (*to*)

- Predicative adjective + *to*-clause: are likely to be, are more likely to, is likely to be, is very important to, must be able to, not be able to, should be able to, will be able to, will not be able to, would be able to, would be difficult to
- (Passive) verb phrase +*to*-clause: are not allowed to, can also be used+(to), can be used to, (can)+have the right to, could be used to, has the right to, is considered to be, seems to have been, should learn how to, want to be a, will be used to, would have to be, would need to be
- *to*-clause: and to be a, in order to achieve, in order to avoid, in order to be, in order to maintain, in order to make, in order to minimise, in order to understand, to be able to, to be added to, to cope with the, to enable them to, to ensure that the, to take into account

H. Passive verb + prepositional phrase fragment (PasPP)

- +be seen as a, be taken into account, be used in the, can be applied to+(the), can be divided into, can be explained by, can be found in, can be regarded as, can be seen as+, can be seen in, can be used for, could be seen as, is based on the, is concerned with the, is illustrated in figure, should be placed on, was followed by a

I. Pronoun/Noun phrase + *BE*/verb phrase (S. +V.)

- *there* + be: but there are still, there are a lot of+, there are quite a+(lot of)+, there are so many, there are still some, there are too many, there is evidence that, there is no evidence, there will be a, there would be no
- Pronoun/noun phrase + BE/verb phrase: an example of this+(is), essay is going to, that need to be, that is to say, the main reason is, the most important thing+(is), this is due to, this is due to+(the), this may be due to, this means that the
- Pronoun + be: all of them are, I am going to, most of them are, some of them are
- Personal pronoun + verb phrase: I think it is, I think that this, I would like to, some people think that+(the), we can say that, we can see that, we can see the

This category originally comes from two LSWE groups: 'Pronoun/noun phrase + *BE* (+...)' from academic prose and 'Personal pronoun + lexical verb phrase (+ complement-clause fragment)' from conversation. Since these two patterns both consist of a subject, including existential *there*, and a predicate, they are combined into one category.

The four-word combination *that is to say* is a peculiar case here, which is a fossilised formulaic expression which appears to be non-compositional in terms of its semantics. Yet here the categorisation is purely established on the superficial structure of a lexical bundle just as other formulaic four-word bundles such as *on the other hand*, *at the same time* are categorised under prepositional phrases (PP-based bundles). The majority of concordance lines indicate that *that is to say* generally functions as an adverbial modifying the sentence/clause that follows it, but there is no other appropriate category to accommodate this peculiar bundle. Along with some PP-based bundles which also function as adverbials (e.g. *on the other hand*), this suggests a potential problem with such a taxonomy established on superficial structures of bundles.

J. Verb phrase with active verb (VPf)

- (played)+an important role in, bear in mind that, become more and more, bring a lot of, has a lot of+, have a lot of+, has a number of, meet the requirement of, pay more attention to, taking into account the, will focus on the

K. Copula *BE* + noun phrase/adjective phrase (Copula *BE*)

- is a kind of, is an example of+(a), is by no means, is more important than, is not only a, is one of the, is the fact that, is the most important, is totally different from, was not so much, was one of the

L. Adverbial clause fragment (Adv clause)

- as I have mentioned, as shown in fig, as we all know, as we have seen, as we shall see, because it is not, because they are not, if there is a

M. Others

- as long as the, as soon as the, as well as the, last but not least, than that of the, whether or not to

The three bundles *as long as the*, *as soon as the*, *as well as the* are problematic as it is found that in the concordance lines they sometimes function as a conjunction and sometimes as part of the comparative pattern '*as* Adverb *as*'. This category is distinctive with another category 'Adverbial clause fragment' in the sense that the bundles under 'Adverbial clause fragment' all contain a verb whereas bundles in this 'Others' category do not. These bundles are also categorised under 'Others' in Biber et al's (1999) framework.

It has to be stressed that the quantitative analysis in this thesis will be presented in the form of an overall tendency of use of NP-based (Categories A-B), PP-based bundles (Categories C-D) and VP-based bundles (Categories E-L), as this is a major distinction between registers discovered in the literature, and hence too fine-grained a hierarchical classification system may not serve this purpose well. The subcategories demarcated here are referred to only in further qualitative discussion following quantitative results. There are two reasons why conducting quantitative analysis on the basis of the subcategories is not considered. On the one hand, the ambiguity across the delineation of subcategories would greatly decrease the validity of such analysis. On the other hand, there are simply not sufficient instances in the subcategories in each corpus for statistical tests to be effectively carried out. The purpose of categorisation addressed above, thus, is used to facilitate the researcher to find the patterns of usage in terms of structural distribution as we shall see in the next two analysis chapters.

4.3 Functional Classification

In the previous section, the problems concerning structural categorisation have been fully discussed. Similar methodological issues also arise when it comes to bundle categorisation in terms of the discourse functions, and the problem of ambiguity in functional classification might be even more complex.

Unlike structural categorisation, which can be mostly determined from analysing the surface grammatical structure within a lexical bundle, functional categorisation has to take into account the context which each instance occurs in. However, some lexical bundles can carry multiple discourse functions within one context while the functions of a number of bundles are context dependent, all of which means that assigning a lexical bundle to merely one functional category might fail to subsume other functions that a bundle could possibly serve. Additionally, sometimes the boundary between different categories can be difficult to demarcate.

This section will start by a review of how functional categorisation has been previously carried out by other researchers. Then the specific problems with respect to categorisation will be illustrated with some examples. Finally, we shall see whether it is possible to improve the current framework established by Biber and his fellow researchers (2003; 2004; 2007) by checking its practicability when applied for analysis.

4.3.1 Background

Biber et al. (2003) first proposed an initial taxonomy for lexical bundles in speech and writing, in which the functions were divided into four major categories: (1) referential bundles, (2) text organisers, (3) stance bundles, and (4) interactional bundles. Each of these categories contained several subcategories, e.g. stance bundles which consist of epistemic, desire, obligation and intention bundles. In a follow-up study (Biber et al., 2004), this

taxonomy was refined with the addition of more subcategories, and a few adjustments were made. On the one hand, the category of interactional bundles seems to have been downsized with the new name of 'Special Conversational' containing only a few lexical bundles. On the other hand, the subcategory of framing (e.g. *in the absence of*) was moved from text organisers to referential expressions. In Biber & Barbieri's most updated taxonomy (2007), the category of interactional bundles or special conversational bundles disappeared, and the subcategory of identification/focus (e.g. *is one of the*) was reassigned from referential expressions to discourse organisers. Table 4-4 presents the categories and subcategories for the development of functional taxonomy described above.

Table 4-4 Development of taxonomy for functional categorisation in studies conducted by Biber et al.

Biber et al. (2003)		**Biber et al. (2004)**		**Biber & Barbieri (2007)**	
Stance Bundles	• Epistemic • Desire • Obligation • Intention	Stance Expressions	• Epistemic • Attitudinal/modality (desire, obligation/directive, intention/prediction, ability)	Stance Expressions	• Epistemic • Desire • Obligation • Intention/ prediction • Ability
Text Organisers	• Contrast/comparison • Inferential • Framing	Discourse Organisers	• Topic introduction/ focus • Topic elaboration/ clarification	Discourse Organisers	• Topic introduction • Topic elaboration/ clarification • Identification/ focus
Referential Bundles	• Time markers • Time/place/text deixis	Referential Expressions	• Identification/focus • Imprecision • Specification of attributes (e.g. quantity, framing) • Time/place/text deixis • Multi-functional reference	Referential Expressions	• Imprecision • Specification of attributes • Time/place/text deixis
Interactional Bundles	• True inquiry • Reporting • Imprecision tags • Politeness markers	Special Conversational	• Politeness • Simple inquiry • Reporting		

* The shifting subcategories are indicated with underlines.

Two major adjustments are most noticeable during the process of taxonomy development in the series of studies conducted by Biber and his fellow researchers. One is the deletion of the category of interactional bundles/special conversational, which could be due to the fact that there are simply not enough bundles of this type to maintain this category and also that most of the conversational bundles actually have a specific discourse function and therefore can be assigned to other appropriate categories. The other notable change is the shifting categorisation of a few certain types of lexical bundles such as framing bundles defined by Biber et al. (2003) (see Figure 4-1), which foregrounds the complexity of functional categorisation.

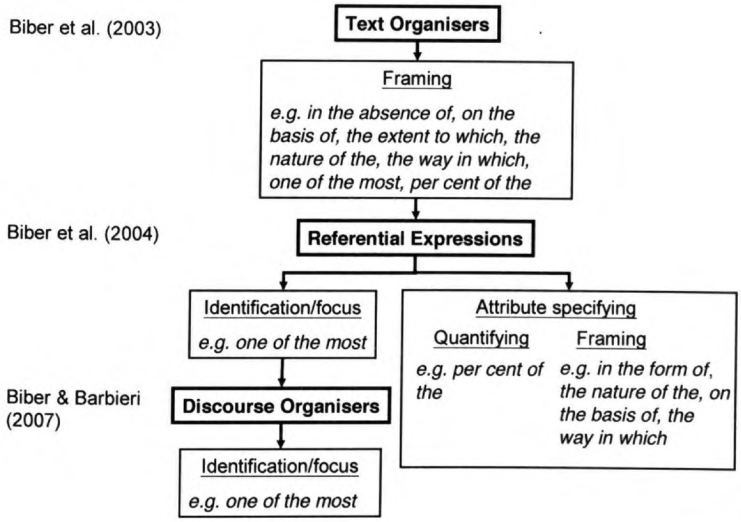


Figure 4-1 Shifting categorisation of certain types of bundles

In Figure 4-1, it can be seen that one broad subcategory, 'framing bundles' in text organisers in the first proposed taxonomy (Biber et al., 2003), was divided into two subcategories later, 'identification/focus' and 'attribute specifying' and moved to the category

of referential expressions (Biber et al., 2004). The finer-grained classification in the second study appears to be more justifiable in the sense that this group of bundles indeed demonstrates diverse specific functions such as quantifying or referring to certain attributes of the entities. The shifting categorisation also reveals that there is absolutely no clear-cut demarcation between categories, particularly for discourse organisers and referential expressions. Even though the identification/focus bundles like *one of the most* are grouped under discourse organisers in the latest study (Biber & Barbieri, 2007), they are still delineated as 'referential identification/focus' in the same study (*ibid*, p.271).

It is also worth mentioning that Hyland (2008a, 2008b) devised another functional taxonomy for the genre of academic writing only. He divided the bundles into three primary categories, each of which concerns one aspect of research writing: research-oriented, text-oriented, and participant-oriented. However, as the current project compares not only academic writing but also learner essays on argumentative or expository topics (termed as EAP-like writing), a neutral framework such as Biber et al.'s (2003) which can be applied on academic and non-academic texts is more suitable for this purpose.

In the following section, the complex issues pertaining to functional categorisation will be discussed further.

4.3.2 Issues Concerning Functional Categorisation

Apart from the ambiguity in the delimitation of functional categories, more problems are also encountered when fresh lexical bundles which have not been covered in the reported studies emerge in the current project, particularly those learner-exclusive bundles. The most noticeable difficulty when applying the classification system is how to determine the category when one word combination has multi-functions in one single context or when it is context dependent. In terms of possible multi-functions, take the majority of bundles in the 'anticipatory *it*' pattern for example. Expressions such as *it is possible to* or *it is difficult to*

can be grouped as stance expressions to express epistemic or ability, but they can also be seen as discourse organisers which identify or focus on the proposition to come. The nature of certain lexical bundles which are composed of more than one functional component might contribute to multiple functions as well, particularly those with modal auxiliary verbs since modal auxiliaries generally carry epistemic or attitudinal/modality senses. The bundle *may be due to* consists of two functional components with *may be* delivering the degree of epistemic certainty and *due to* expressing inferential assumption. Another example *it should be noted* exhibits a similar dilemma of whether the bundle should be classified as an obligation/directive stance or as an identifying discourse organiser. What makes this worse is that epistemic and deontic modality is not always easy to distinguish (e.g. *must be able to*, *would have to be*, *would need to be*).

In terms of context-dependent functions, an examination of concordance lines shows that a bundle does not always have the same function in each context. Sometimes the concordance lines retrieved from one single lexical bundle could occur in different contexts and thus have different functions. A complicated instance of this found in BAWE-EN is *to a lack of*, in which the preposition *to* forms part of a range of expressions *due to*, *down to*, *in response to*, and *lead to* (see Table 4-5 for the word combinations and frequency). This example also involves two functional components in one word combination just discussed above. The majority of the first components, *due to*, *down to*, and *lead to*, express a sort of inferential connotation. The second component *a lack of* contributes to the discourse function of reference as well.

Table 4-5 Expressions incorporating *to a lack of* and their frequency in BAWE-EN

Expressions	Freq
due to a lack of	2
down to a lack of	1
in response to a lack of	1
lead to a lack of	1

It appears that the rule of thumb to deal with the task of assigning a bundle to an appropriate functional category is to go through the concordance lines and give priority to 'the most common use' (Biber et al., 2004, p.384). As a result, the process of determining a corresponding category turns out to be extremely time-consuming as the concordances have to be examined in great detail whenever in doubt. The whole procedure for functional classification will be described in the next section.

4.3.3 A Modified System for Functional Classification

The taxonomy established by Biber and his colleagues does not offer an all-round solution to the issue of functional overlapping and ambiguity, but it provides a good basis for the classification to be carried out. Basically, the most updated classification used in Biber and Barbieri (2007) was adopted with some necessary modification to better suit the written data investigated in this project. Some subcategories, such as 'imprecision referential' (e.g. *or something like that*), 'desire' (e.g. *if you want to*), or 'intention/prediction' (e.g. *is going to be*), were removed from the categorisation as no bundles fit the description and they all appear to relate to spoken data. In addition, a couple of subcategories under stance expressions—'obligation/directive' and 'ability'—were combined as 'attitudinal/modality' just as in Biber et al. (2004), since there are not enough instances to sustain these subcategories. On the other hand, because too many bundles meet the requirement of the referential subcategory 'specifying attributes', they are divided into two groups: 'framing' and 'quantifying'.

In the process of categorisation, several steps have been taken with an attempt to refine the system. The first step was to revisit the definition of each subcategory and allocate the lexical bundles retrieved in this project to the categories which are considered to best accommodate them. The assigned category in this project was then compared with the literature whenever a bundle in question was reported in any of the studies which adopted the same or a similar framework (*cf.* Biber et al., 2003; Biber et al. 2004; Biber and Barbieri,

2007; Cortes, 2002; Cortes, 2004). Any disagreed category allocations were highlighted and re-examined. After the reviewing process, a number of lexical bundles have hence been repositioned to corresponding categories which are thought to be more appropriate than the categorisation reported in the literature. Bundles such as *the extent to which* or *to a large extent*, for instance, are considered to specify the attribute of degree and extent and thus are more suitable to fall within the subcategory of quantity specification rather than that of intangible framing attributes. Another instance is *as a result of*. It was assigned to the intangible framing referential expressions category in Biber et al. (2004) but has been repositioned in the category of inferential text organisers on the ground that this expression is generally considered to connect the prior and forthcoming texts by conveying the relationship of cause and effect. Sometimes the function of a bundle utterly depends on the context it resides in. Such instances occur most in deictic referential expressions such as *the end of the* or *at the beginning of* because whether they are place deictic or time deictic is up to the context. In order to simplify the system, it was decided not to distinguish those deictic referential expressions although the individual deictic functions were annotated in the working databank. However, for other ambiguous cases, the concordance lines have to be re-checked and the frequency counts of each function a bundle has in the concordance lines has to be recorded. Therefore, the second step was to examine the concordance listings and take notes of the discourse functions that each bundle can possibly have. Some contextual information, such as the words prior to and following the target bundle and the corresponding frequency counts, was also documented for the sake of category determination. Special attention was paid to cases when one bundle overlaps with another one or two. This practice attempts to solve two problems mentioned earlier. On the one hand, some bundles may seem to have more than one discourse function from the inspection of their surface structure. For example, *as we have seen* can be a discourse organiser elaborating a topic or a referential

expression indicating the text deixis. Checking concordance lines helps determine its primary function as being text deictic. On the other hand, the contextual information recorded could be used to solve the issue of overlapping bundles, which has been discussed in the previous chapter.

The finalised functional categorisation contains three major categories: referential bundles, stance bundles, and discourse organisers, and each of them can be divided into a number of subcategories to accommodate the lexical bundles under investigation in this project, as have been described in the beginning of this section.

Referential Expressions

The referential category is characterised with the function of attribute specification. The first type, framing bundles, are used to specify a given attribute or condition and typically composed of noun phrase fragments or prepositional phrase fragments. Another common type of referential bundle is quantifying expressions that qualify the proposition with expressions related to anything potentially gaugeable such as size, number, amount, extent, etc. The last subcategory of referential expressions is place/time/text-deictic bundles.

A. Framing bundles:

- *Here we organize the methods and results of this literature in the context of the very diverse arrangements observed in European countries.* (FLOB-J)
- *The issue of different time motivations stems from the nature of the work that each department does and would be difficult to change.* (BAWE-EN)
- *Although the common market had not been fully developed, a wide range of policies which affected the relationship between the nation state and the community were in place.* (CEFR-C1)

B. Quantifying bundles:

- *In part this stemmed from the growing general governmental involvement in a wide range of external economic security policy.* (FLOB-J)
- *However, to argue this is to ignore the extent to which the politics and language of Empire entered into feminist discourse.* (BAWE-EN)
- *However, if we take a look at some other parts of the world, we could notice that there are a lot of problems facing the world that we are living on today.* (CEFR-B2)

C. Place/time/text-deixis bundles:

- *Example time series are shown in Fig. 9.3.* (FLOB-J)
- *In the Great Britain, voluntary restrictions on both DDT and dieldrin started in 1962, and completed at the end of 1975.* (BAWE-CH)
- *Traffic congestion occurs for a long time in Tokyo. The roads are mostly designed for lighter traffic of several decades ago.* (CEFR-C1)

Stance Bundles

Stance bundles can be used to express the writer's epistemic evaluation of a proposition in terms of its certainty or uncertainty. A large number of epistemic bundles are hedge devices, used to mitigate the extent of impact in the proposition, e.g. *is likely to be* or *seems to have been*. Many of the epistemic bundles can also serve as discourse organisers to identify or elaborate the subject matter, e.g. *to the fact that* or *it has been suggested*. With respect to the form, the commonest structure in epistemic bundles is the 'anticipatory *it*' pattern. See the following illustrated examples extracted from the corpora investigated:

D. Epistemic bundles:

- *A more complex case occurs when some conditions are more likely to produce missing data than others but within each condition each observation is equally likely to be missing.* (FLOB-J)

- *A potential flaw with the EIA process in the UK is the fact that the ES is produced by the developer.* (BAWE-EN)
- *Some people think that laws allowing abortion will increase irresponsible pregnancies and lead to disrespect for human life.* (CEFR-B2)

Stance bundles can also convey the writer's attitude about the forthcoming proposition. As mentioned earlier, this subcategory, attitudinal/modality, combines sub-functions such as obligation/directive and ability. The former informs what the writer thinks is obligatory or directs the reader for some action while the latter generally involves the writer's judgment on the capability of doing something.

E. Attitudinal/modality:

- *However, they argue that it is difficult to explain the wage differences, for instance, of secretaries across industries with the same model.* (FLOB-J)
- *It is necessary to provide for unpredictable complex social interactions in captivity, and allow individual's to have a choice over companionship.* (BAWE-EN)
- *Therefore, the "first generation" model needs to be extended in order to be able to explain these findings.* (BAWE-CH)

Discourse Organisers

Discourse organisers are used to structure texts. They can introduce or elaborate a topic and make inference. In addition, a large number of discourse organisers discovered here function to identify the focus that the writer stresses, usually a noun phrase fragment (e.g. *one of the most*) or a clause fragment (e.g. *that there is a, we can see that*) referring to an entity or a proposition. This type of identification/focusing bundles sometimes may also be grouped into referential bundles if they are interpreted as making reference to an entity.

F. Topic introduction bundles:

- *Facing the trend, this essay is going to explore tourism distribution channels on the Internet.* (BAWE-CH)

G. Topic elaboration bundles:

- *This last point is described in more detail in the following section on noise and flexibility of changing strategy.* (FLOB-J)
- *One normally needs to attend formal lessons in order to acquire a second language. On the other hand, formal lessons are not required in the acquisition of their first language, neither is any learning aids required.* (CEFR-C1)

H. Inferential bundles:

- *Even the fact that such county-based regionalisation does not allow division between rural and urban areas can be tolerated in view of the compensating advantage this classification allows in long-term internal consistency of the data.* (FLOB-J)
- *People in Hong Kong are facing 1997 which is the time when china Government will come and make Hong Kong communist. As the result of this, many people are immigrating to other countries and the future of Hong Kong is still very difficult to tell.* (CEFR-B2)

I. Identification/focusing bundles:

- *This is perhaps one of the most significant limitations of the Tiebout model.* (FLOB-J)
- *In this case, there would be no charge between the firms, $t = 0$.* (BAWE-EN)
- *As far as the children are concerned, learning some skills does not mean serious training.* (CEFR-C1)

4.4 Conclusion

This chapter began with a description of how the comparisons of bundles in different corpora would be carried out in the following two analysis chapters. Firstly, two modular studies were conceived in light of the different genres being examined. The first modular study compares native and non-native writing in an academic context while the second modular study mainly deals with argumentative and expository learner writing between two proficiency levels. Secondly, bundle types and tokens are distinguished in order to reflect the range and intensity of the use of recurrent phraseology. Then the keyness analysis would be introduced to complement the bundle analysis which is established on a frequency and dispersion threshold.

This chapter also described in detail how the lexical bundles were categorised on the basis of structures and discourse functions. As can be seen, the current structural and functional categorisation adopted from the LSWE taxonomy required some modification in order to accommodate the bundles retrieved from the corpus data in this thesis, particularly the learner idiosyncratic ones. In addition, the nature of such an automated approach also led to several problems in categorising the bundles. Consistency, however, is one of the keys to the success of category annotation, despite the fact that the documenting process is fairly tedious and arduous, yet great efforts have been made to ensure that the categorisation conforms to all the definitions and specifications.

In the next chapter, the analysis of Modular Study 1 built upon the rationale addressed in this chapter will be illustrated.

Chapter 5 Modular Study 1: Lexical Bundles in Native Writing and Learner Writing within the Academic Context

In this chapter, academic writing in three groups (native academicians, native students, and non-native students) are investigated and compared in terms of their use of lexical bundles. First, the process of data selection is described. In addition, some crucial linguistic information regarding the three selected groups of writing are presented so as to provide a background for the comparisons of lexical bundles made later. Then the structural, functional and keyness analyses, as summarised in the previous chapter, are conducted. The relationship between the structural and functional analyses is also discussed. This chapter will finish with a summary of major findings. Overall, this chapter aims to address the second and third analytical research questions (see Section 3.1) by revealing how the use of lexical bundles in learner performance differs from native language in terms of structures and functions by various measures.

5.1 Selection of Data

Two existing corpora were used for the present study: the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus and the British Academic Written English (BAWE) corpus, both of which cover a wide range of disciplines (For BAWE, see Alsop & Nesi, 2009; for FLOB, see Hundt, Sand, & Siemund, 1998). To ensure comparability, only part of each corpus was selected for investigation. The FLOB corpus is a one-million-word corpus of written British English from the early 1990s, containing fifteen genre categories. For the current study, only the category of academic prose, FLOB-J, was used as the representative group of native expert writing so as to mirror student writing within the academic discourse. FLOB-J is composed of eighty 2,000-word excerpts from published academic texts retrieved from journals or book sections.

With regard to the L1 and L2 student academic writing, parts of the BAWE corpus were utilised. The BAWE corpus contains approximately 3,000 pieces of proficient assessed student writing, which amount to 6.5 million words in total. Holdings are widely distributed across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences), thirty-five disciplines, and across four levels of study (typically three years in undergraduate studies and one year in the taught Masters level). As some students contributed more than one assignment to the BAWE corpus, it was decided to adopt only one piece of written sample from each student in order to guard against idiosyncrasies of individual writing style. Two student subcorpora were selected from the BAWE corpus: one is BAWE-CH, which contains essays produced by L1 Chinese students of L2 English, while the other, BAWE-EN, is a comparable dataset contributed by peer L1 English students.

Students' backgrounds are well documented in the BAWE corpus, which allowed the researcher to examine closely whether L2 learners had received secondary education in the UK or not and if so how long the length of British secondary education lasted. Given that the size of target learner essays contributed by those who never received any British secondary education is still much smaller in comparison with BAWE-EN and FLOB-J, it was decided to also include the texts written by those having received British secondary education for no longer than two years to increase the corpus size (see Table 5.1). This decision is considered acceptable as L2 learners who have only received a couple of years of secondary education in the UK are unlikely to have reached native-like proficiency. On the other hand, some of the L1 Chinese learners of L2 English who had received British secondary education for a substantial period of time, say five years or longer, might have reached a certain level of native-likeness in their L2 English. It is therefore sensible to remove any essays contributed by those possibly more native-like learners as they might undermine the quality of learner data when compared with native peer writing. Indeed, the determination of learners' L1 and

L2 backgrounds might seem arbitrary in some sense, but at least this extra information helps better control the data quality than most of the reported second language research. The decisions above, including using only one piece of writing from each student, reduced the learner data BAWE-CH from originally around 400,000 words to nearly 150,000 words. This size, however, is relatively comparable to FLOB-J and BAWE-EN as we shall see later.

Table 5-1 Number of BAWE samples from L1 Chinese students who received different years of secondary education in the UK

Years	*UK 0*	*UK 0.5*	*UK 1*	*UK 1.5*	*UK 2*	**Total**
No of samples	43	1	3	1	5	53

Another group of student writing extracted from the BAWE corpus is native peer writing, BAWE-EN, which is the L1 English counterpart of L2 learner data in the current study. Due to the large amount of native data in BAWE, written samples of native students were granted the privilege of being carefully selected, taking into account of year of study, grade (merit or distinction), genre[20], and disciplinary area. Great efforts were made to ensure the native samples were widely distributed in terms of those variables (see Table 5-2, Table 5-3, and Table 5-4 for comparisons of BAWE-CH and BAWE-EN).

[20] The variable of genre might not be a reliable indicator of text type as this variable was identified by the contributors within the choices of *case-study*; *essay*; *exercise*; *notes*; *presentation*; *report*; *review*; and *specified other*, and this self-provided information often failed to match the labels students themselves used metatextually. For example, an assignment labelled as an *essay* might begin with 'In this report', or vice versa (Alsop & Nesi, 2009, p. 76). This information, therefore, is mainly used for supplementary purposes rather than being definitive.

Table 5-2 Number of samples written by students from different level of study in BAWE-CH and BAWE-EN

Corpus	Level of study (Year 1 to Year 4/Masters)	
BAWE-CH	Y 1	13
	Y 2	5
	Y 3/4*	10
	Y 4/M	25
	Total	*53*
BAWE-EN	Y 1	15
	Y 2	15
	Y 3/4	15
	Y 4/M	15
	Total	*60*

*Most undergraduate courses in British universities last for three years, but there are also four-year undergraduate courses. In such cases, the fourth year is sometimes part of undergraduate studies and otherwise at the Masters level. Assignments written in Year 4 were thus categorised as level three (in the case of Year 3 being an intercalatory year) or level four (Masters level) on the basis of information provided by the student contributors.

Table 5-3 Number of samples awarded the grade of merit (M) or distinction (D) in BAWE-CH and BAWE-EN

Corpus	Grade (Merit/Distinction)	
BAWE-CH	M	27
	D	23
	Unknown	3
	Total	*53*
BAWE-EN	M	34
	D	26
	Total	*60*

Table 5-4 Number of samples from different genres in BAWE-CH and BAWE-EN

Corpus	*Genre*	
BAWE-CH	case study	5
	critique	8
	essay	20
	proposal	3
	methodology recount	7
	others	10
	Total	*53*
BAWE-EN	case study	3
	critique	8
	essay	33
	proposal	3
	methodology recount	5
	others	8
	Total	*60*

The texts in FLOB-J were also categorised by use of the same four disciplinary areas defined in BAWE: Arts and Humanities (AH), Life Sciences (LS), Physical Sciences (PS), and Social Sciences (SS). The written samples in the three corpora are quite broadly distributed, with the only exception of merely one learner essay in BAWE-CH falling into the area of Arts and Humanities (see Table 5-5).

Table 5-5 Number of samples from different disciplinary areas (Arts and Humanities (AH), Life Sciences (LS), Physical Sciences (PS), Social Sciences (SS))

Corpus	Disciplinary areas	
BAWE-CH	AH	1
	LS	15
	PS	12
	SS	25
	Total	*53*
BAWE-EN	AH	15
	LS	15
	PS	15
	SS	15
	Total	*60*
FLOB-J	AH	22
	LS	15
	PS	21
	SS	22
	Total	*80*

The sizes of the three finalised corpora are considered to be fairly comparable, as presented in Table 5-6, the average of which is around 150,000 words. At first sight, the corpora used might seem fairly small, particularly in comparison with most L1 corpus-based studies. In the context of second language research or learner corpus studies, the present corpus size, however, appears to be comparably sufficient (cf. individual subcorpora of 100,000-200,000 words by learners of different L1s in the International Corpus of Learner English (ICLE)). Meanwhile, the three corpora for comparison have been carefully designed and matched, which hence should help to mitigate any negative effects of the overall small size. From the pilot categorisation reported in Section 4.2, the structural distribution of bundles retrieved from FLOB-J revealed a surprisingly similar pattern with those extracted from the 5.3-million-word academic prose in the LSWE (Longman Spoken and Written English) corpus, which suggests that even the corpora as small as the current investigated ones can still be used to effectively retrieve representative recurrent phraseology.

Table 5-6 Constituents of the three academic corpora

Representation	*Corpus*	*Word count*	*Average length of text*	*No of texts*
Learner writing	BAWE-CH	146,872	2,771	53
Native peer writing	BAWE-EN	155,781	2,596	60
Native professional writing	FLOB-J	164,742	2,059	80

With respect to L1 Chinese learners' backgrounds discussed in Chapter 2, although the BAWE corpus documented the students' mother tongues, however, it did not specify which variety of Chinese language each contributor's L1 was (e.g. Cantonese, Mandarin, or Taiwanese). Instead, only the umbrella term, Chinese, was recorded. Neither was the place of origin of these L2 students documented either (e.g. Hong Kong, China, Taiwan, Singapore, etc.). However, it would be reasonable to assume that these L2 students came from heterogeneous backgrounds which more or less reflected the constituents of L1 Chinese students in current British higher education rather than from a homogeneous community.

Although the three groups, FLOB-J, BAWE-EN, and BAWE-CH, have been considered to be fairly matched in terms of text types or genres, some doubts might still arise as to the difference between published academic texts and university student essays. One potential problem emerging is that FLOB-J is composed of excerpts of book sections and journal papers whereas BAWE contains complete texts only. However, in the results of analysis, we will see that the impact made upon the lexical bundles is fairly minimal.

5.2 Linguistic Profile

This section is intended to provide some background information in regard to the three written subcorpora investigated so that an overall picture of the groups of texts to be compared can be presented. The first section addresses the issues of type/token ratio, a lexical measure which has been extensively used in both L1 and L2 research. The next section shows

the number of words with various lengths in the three groups of writing investigated. Then some examples of the titles or topics are illustrated in the final section.

5.2.1 Type/token Ratio

The type/token ratio (TTR) is a very common measure used to investigate lexical complexity and variation in language development studies (Wolfe-Quintero et al., 1998, p. 101). The higher the type/token ratio is, the richer the writer's lexicon is considered to be. *WordSmith 4.0* provides a function called the standardised type/token ratio (STTR), which computes TTR every 1,000 words, so as to eschew the criticism that the TTR measure is constrained by text length.[21] As can be seen from Table 5-7, all three subcorpora have very similar STTRs (at around 39), although both sets of student writing have slightly lower STTRs than the writers in FLOB-J, and this ratio in BAWE-EN is slightly lower than that in BAWE-CH (39.16% vs. 39.3%).

Table 5-7 STTR in Modular Study 1

	BAWE-CH	***BAWE-EN***	***FLOB-J***
STTR (per 1,000 words)	39.3	39.16	39.64

It is not surprising that published academic texts shows the highest STTR since professional academics are supposed to be the most proficient writers among the three groups and would be expected to have a wide range of lexis (we might have expected FLOB-J to have an even higher STTR). It is, however, surprising to see that British student writing appear to be nearly identical with Chinese student writing in terms of lexical variation as

[21] TTR partly hinges on text length. Longer texts are very likely to have lower TTR than shorter texts because many grammatical words (*the*, *of*, *or* etc.) would unavoidably reoccur in a longer text. The TTR measure has hence been long criticised, and many modified measures based on TTR have accordingly been developed.

English is their mother tongue, which could be a result of various factors. Perhaps Chinese students resorted to dictionaries or thesauri more often than British students did when encountering certain concepts they could not express well with the target language. Some learners may thus have 'borrowed' novel or complex-looking lexical items which are not in their mental lexicons. Or the Chinese STTR might be attributed to a common myth among L2 student writers that demonstrating their knowledge of lexis can add more weight to the quality of their writing, and they thus purported to use longer and more complicated vocabulary.[22] The above assumptions could be further supported in the following section. However, on the whole, the STTRs are too similar to point to any statistically significant differences – another explanation could be that there is hardly any difference in lexical variation due to genre constraints once a writer reaches a certain standard.

5.2.2 Words with Different Lengths

The wordlists were generated by *WordSmith*, which indicate the numbers of words with different lengths (words of different numbers of letters). In Table 5-8, it can be seen that the Chinese students have largely demonstrated fewer occurrences of words than the other two groups of native writers from one-letter words to eight-letter words. Even if there was no difference between the abilities of writers in the three sets of data, this might be expected, considering that the Chinese data is slightly smaller than the other two subcorpora. Nevertheless, starting from the row for nine-letter words, the numbers in the column representing BAWE-CH begin to exceed the numbers in BAWE-EN and sometimes even those in FLOB-J. This is not what would have been expected in the sense that L2 writers are supposed to be the least competent writers in the three groups and thus should have shown

[22] These assumptions took shape from the researcher's observations on L1 Chinese students' writing in the UK and also from the researcher's own experiences as a L2 speaker of English.

the least lexical complexity and variation. Such kind of comparison based on word length, to my knowledge, has not been considered in second language research, and whether the above assumption holds true certainly still needs more empirical evidence, but it could be an interesting topic for further research.

Table 5-8 Numbers of words with different lengths in Modular Study 1 (occurrences)

N	BAWE-CH	BAWE-EN	FLOB-J
1-letter words	5406	5718	6690
2-letter words	24926	27640	30050
3-letter words	26326	28031	29407
4-letter words	20228	23453	22781
5-letter words	14258	14880	16042
6-letter words	12419	12408	12482
7-letter words	12381	12715	13072
8-letter words	10610	10677	11014
9-letter words	8529	7655	8386
10-letter words	5958	5437	6204
11-letter words	4056	3473	4016
12-letter words	2068	1808	2117
13-letter words	1362	1194	1263
14-letter words	551	440	511
15-letter words	239	136	242
16-letter words	61	46	111
17-letter words	50	53	30
18-letter words	31	20	22
19-letter words	17	12	11
20-letter words	10	10	7

5.2.3 Topics & Titles of the Written Samples

To better reflect the nature of groups of writing compared, a random selection of essay/paper topics and book/chapter/paper titles included in the three subcorpora are presented in this section under the four broad disciplinary areas (Arts and Humanities (AH), Life Sciences (LS), Physical Sciences (PS), Social Sciences (SS)) as defined by the BAWE annotation guidelines and addressed in Section 5.1.

Table 5-9 Some examples of essay/paper topics and book/chapter/paper titles in Modular Study 1

Area	BAWE-CH	BAWE-EN	FLOB-J
AH	• A Review of Factors Affecting Degree of L2 Foreign Accent*	• Why did Britain lose the American Revolutionary War? • Why is the Carolingian period so important for the transmission of Latin texts? • Which theoretical approach has best helped you 'make sense' of The Waste Land and why?	• Language, Thought, and Falsehood in Ancient Greek Philosophy • Sir Edmonds and the Official History: France and Belgium • Gender and Narrative in the Fiction of Aphra Behn
LS	• Monoclonal Antibody Structure and Function Determination with Biophysical Techniques • Can the food industry be blamed for the increase in obesity in the UK? • Loss of seed quality during storage	• HIV/AIDS and stigma • Conserving Wild Mammal Species in the Farming Environment • Discuss the implications of herbicide revocation following implementation of EU directive 91/414 on UK horticulture.	• Infectious Diseases of Humans • Cognitive Development: An Information Processing Approach • Land Degradation: Development and Breakdown of Terrestrial Environments
PS	• Humanoid Robotics in Artificial Intelligence • Assignment 2: A stylus type instrument (Profilometer) • Multivariate Statistics Assignment One	• The Mathematics of RSA Cryptography • A kinetic study of the hydrolysis of crystal violet. • The Bohr Model of the Atom	• Mathematics and the Image of Reason • Networks and Telecommunications. Design and Operation • An Introduction to Grain Boundary Fracture in Metals
SS	• Malthusian Trap and Economic Growth • Evaluate the case for reform of Britain's law on industrial conflicts in the light of the Gate Gourmet dispute during the summer of 2005. • In the Search of Political Leaders - A Review of the Bureaucratic Governance of Hong Kong under the Political Thoughts of Max Weber	• Report on the implementation of the Uncrc in the Russian federation. • What were the major changes in the international economy after 1914? Why did the Gold Standard work well before 1914 but not in the interwar period? • Compare the 'functionalisms' of Malinowski and Radcliffe-Brown	• Rational Choice and Politics • Globalization and Global Localization: Explaining Trends in Japanese Foreign Manufacturing Investment • Vocational Qualifications in Britain and Europe: Theory and Practice

* There is only one learner essay in the category of Arts and Humanity in BAWE-CH.

Indeed, it has to be acknowledged that student writing produced at undergraduate and postgraduate levels and the published academic texts simply represent different genres and

thus may not be perfectly comparable. From the broader scope, however, these texts were all generated under academic settings. A quick overview of the topics or titles from the text corpora investigated provides a gist of what sort of texts are compared before we move on to the lexical bundle analysis in the rest of this chapter.

5.3 Structural Analysis

Two kinds of structural analysis were carried out in this section. The first analysis reveals the *type* distribution of lexical bundles, and the second demonstrates the *token* distribution of them. As discussed earlier, it is believed that by taking into account the distinction between types and tokens, the extent of difference in the use of lexical bundles among the three corpora can be more thoroughly reflected because it is likely that a corpus could exhibit a narrow range of lexical bundles yet with very high frequencies of the bundles, or vice versa.

Six bundle types which do not fit into any of the 'NP-based', 'PP-based', or 'VP-based' categories have been allocated to the category of 'Others' (cf. Section 4.2). These bundles had to be excluded from the significance tests because a chi-square test would not be valid if any of the possible categories entails fewer than five instances in each corpus. These bundles are *as long as the*, *as soon as the*, *as well as the*, *last but not least*, *than that of the*, and *whether or not to* (together they account for only 2.1% of the total 292 types). The first four bundles appear in BAWE-CH while *as well as the* and *than that of the* are used in BAWE-EN. Finally, *whether or not to* is retrieved only from FLOB-J. For the running of chi-square tests, the significance level is set at 0.05 throughout this thesis. If there is a relationship between corpora and structural distribution of bundles ($p<0.05$), the standardised residuals for each cell in the contingency table would be calculated in search for the major contributors to the significant difference. The procedure and rationale of calculating such

residuals will be discussed after the results of structural distribution of bundle types and tokens are presented.[23]

5.3.1 Type Distribution

As can be seen from Table 5-10 and Figure 5-1, professional writing in FLOB-J is quite evenly distributed with nominal phrase fragments (30.8%), prepositional phrase fragments (37.4%), and verbal phrase fragments (31.8%) whereas in student writing represented in BAWE-EN and BAWE-CH, almost half of the bundles are made up of VP-based expressions (54.9% and 50% respectively) and the reliance on NP-based bundles (nearly 16% for each) is much less than FLOB-J.

Table 5-10 Structural distribution in Modular Study 1 (types)

χ^2=14.9, df=4, p=0.005			Structure			Total
			NP-based	PP-based	VP-based	
Corpus	BAWE-CH	Count	12	26	38	76
		% within Corpus	15.8%	34.2%	50.0%	100.0%
	BAWE-EN	Count	16	30	56	102
		% within Corpus	15.7%	29.4%	54.9%	100.0%
	FLOB-J	Count	33	40	34	107
		% within Corpus	30.8%	37.4%	31.8%	100.0%

[23] The Bonferroni Correction was considered in the beginning for multiple comparisons made in this thesis. The significance level set at 0.05 means that there is a chance of 5% of error. In other words, every 1 in 20 chi-square tests might have an error in the significance results in the case of multiple comparisons. The notion of Bonferroni correction (cf. Gries, 2003, p. 82; Meyer, 2002, p. 129) was developed to guard against such a potential risk in spurious positives by further lowering the significance level (i.e. dividing the set p value by times of comparisons to be made with the data). In the present thesis, however, chi-square tests are executed with the cases of bundles in terms of structural categorisation and in terms of functional categorisation, which can be treated as separate sets of data although they both originated from the same corpora. In addition, the type and token comparisons, again, can also be regarded as independent data as they deal with different notions of bundles. On the other hand, the chi-square statistics calculated in this chapter, as will be seen later, all have a p value which is far below 0.05 (0.005 as the highest one). It was therefore decided not to further complicate the statistical analysis by adding the use of Bonferroni correction. Instead, the intention here is simply to provide some explanation as to why Bonferroni correction is not adopted.

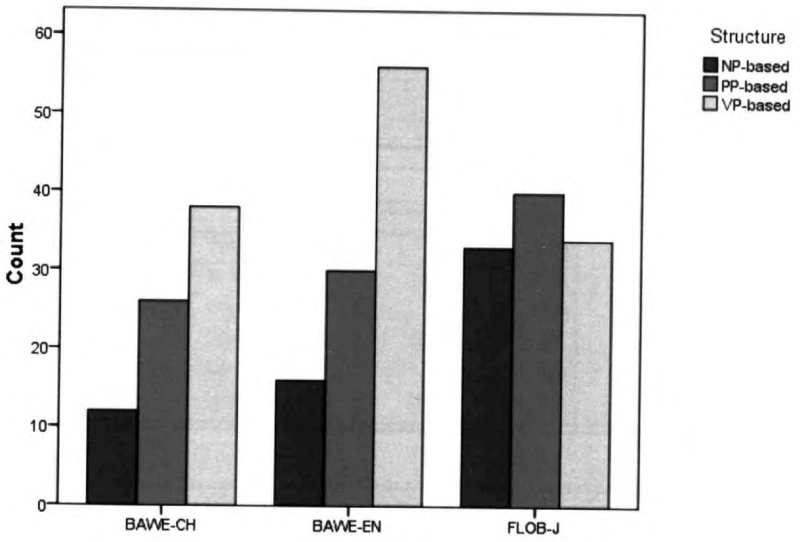


Figure 5-1 Structural distribution of lexical bundles in Modular Study 1 (types)

A chi-square test shows that there is significant difference in terms of structural distribution of bundle types among BAWE-CH, BAWE-EN, and FLOB-J (χ^2=14.9, df=4, p=0.005). As there is significant difference, the standardised residuals are calculated and discussed later.

5.3.2 Token Distribution

The structural distribution of bundle tokens among the three corpora is slightly different from that of bundle types, but the pattern of distribution is still similar for the most part. The proportion of VP-based bundles in FLOB-J falls to 28.6%, yet remains the lowest among the three groups of writers. BAWE-CH appears to be the group which changes most dramatically, with the proportion of PP-based bundles rising to 43.2% and the proportion of VP-based bundles dropping to 43.8%. However, VP-based bundles are still the category that holds the most occurrences in either BAWE-CH or BAWE-EN, and the reliance on NP-based bundles

still remain the lowest in comparison with PP-based and VP-based bundles (see Table 5-11 and Figure 5-2).

Table 5-11 Structural distribution in Modular Study 1 (tokens)

χ^2=103.0, df=4, p<0.0005			Structure			Total
			NP-based	PP-based	VP-based	
Corpus	BAWE-CH	Count	62	206	209	477
		% within Corpus	13.0%	43.2%	43.8%	100.0%
	BAWE-EN	Count	105	205	340	650
		% within Corpus	16.2%	31.5%	52.3%	100.0%
	FLOB-J	Count	199	300	200	699
		% within Corpus	28.5%	42.9%	28.6%	100.0%

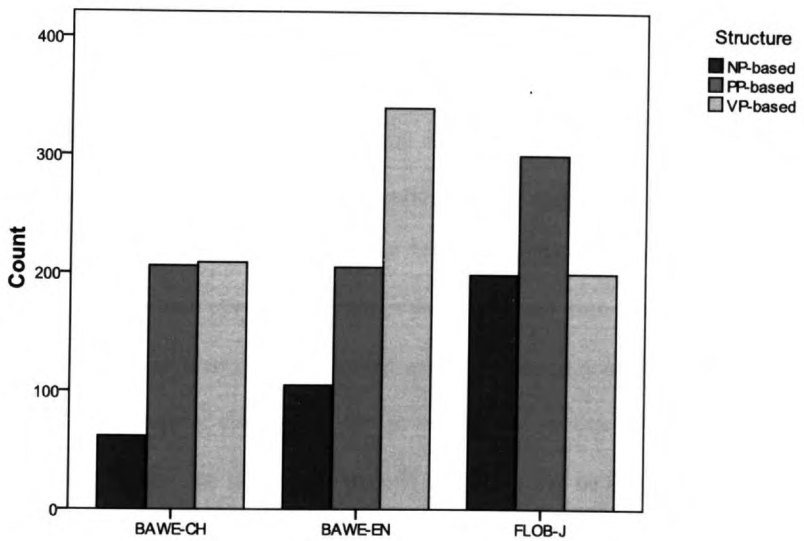


Figure 5-2 Structural distribution in Modular Study 1 (tokens)

A chi-square test indicates that there is significant difference in terms of structural distribution of bundle tokens among the three groups of writing (χ^2=103.0, df=4, p<0.0005). Again, the standardised residuals are calculated and discussed in the next section.

5.3.3 The Chi-square Test and Standardised Residuals

This part of analysis aims to find out that in the case of significant difference found in the distribution of bundle structures between corpora, which structural categories serve to better distinguish the three corpora in terms of language proficiency or, more specifically, writing competency in the current study. At the very beginning, a number of statistical measures, such as the proportion z test or multiple regression, were considered for this purpose. After consulting with a couple of statisticians, it was then decided to stick to the nonparametric chi-square test by calculating its standardised residuals after running each test, which would offer valuable information in this regard.

In this section, the chi-square test is again executed with the three corpora, but this time the focus would be cast on which cells in the three (corpus) by three (structural category) table contribute the most to the significant difference among the three groups of writing so that they can be interpreted as better candidates for distinguishing writing competency. For this purpose, we need to understand first how a chi-square test is computed. A chi-square statistic χ^2 is the sum over all elements in the matrix, and each of the elements refers to the squared difference between the observed and the expected counts divided by the expected count. In a chi-square distribution, the greater the test statistic χ^2 is, the lower the critical value p will be (i.e. the more likely the null hypothesis can be rejected). The formula of the chi-square statistic χ^2 is shown as below:

Equation 5-1 Formula of chi-square statistic χ^2

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where

Oi = an observed count;

Ei= an expected (theoretical) count, if the variable of corpus (representing writing

competency) has no impact on the use of lexical bundles as asserted by the null hypothesis;

n = the number of possible outcomes of each event.

The difference between an observed count and an expected count is called a *residual*. A *standardised residual* (R) allows the researcher to compare more than one contingency table in relation to the overall extent of difference between observed and expected frequencies (Sheskin, 2004, pp. 525-526). By comparing the standardised residuals in the nine cells in a 3×3 table (corpus×structural category) in the current matrix of structural distribution, it can be found out which cells contribute significantly in the chi-square statistics χ^2. A standardised residual R is the residual divided by an estimate of its standard deviation (i.e. the value of an observed count minus an expected count and then divided by the square root of the expected count). The formula is as below:

Equation 5-2 Formula of a standardised residual

$$R = \frac{(Oi - Ei)}{\sqrt{Ei}}$$

By presenting the values of R in a contingency table, we can make a cell-by-cell comparison between each pair of observed and expected counts. A positive value of R indicates the observed count is higher than the expected one, and a negative value of R indicates otherwise. If the absolute value of a standardised residual R for a given cell is greater than 1.96, it suggests that the cell makes a major contribution to rejecting the null hypothesis ($p<0.05$). In the case of structural distribution of bundle types between BAWE-CH, BAWE-EN and FLOB-J, we have already seen that there is a significant difference in the distribution. The standardised residuals are thus calculated to ascertain which cells make a major contribution to the significant difference. As can be seen from Table 5-12, only two cells, the NP-based and VP-based bundles in FLOB-J, have the absolute value of R greater

than 1.96, which suggests that these two categories in FLOB-J make a significant contribution to rejecting the null hypothesis. As discussed above, we can use this information to conclude that native academics used significantly more NP-based bundles and fewer VP-based bundles than expected.

Table 5-12 Chi-square standardised residuals for structural distribution (types) in Modular Study 1

$\chi^2=14.9$, $df=4$, $p=0.005$		NP	PP	VP
BAWE-CH	Observed Count	12	26	38
	Expected Count	16.3	25.6	34.1
	R	-1.1	0.1	0.7
BAWE-EN	Observed Count	16	30	56
	Expected Count	21.8	34.4	45.8
	R	-1.2	-0.7	1.5
FLOB-J	Observed Count	33	40	34
	Expected Count	22.9	36.0	48.1
	R	**2.1**	0.7	**-2.0**

In terms of structural distribution of bundle tokens, as can be seen from Table 5-13, in the 3 by 3 chi-square contingency table, it is still mainly the NP-based and VP-based categories from FLOB-J that make a major contribution to the significant difference. Moreover, every category from BAWE-EN and the category of NP-based bundles from BAWE-CH also contribute to the significant difference. This finding not only further confirms the result from the statistical analysis of type distribution but also suggests that both student groups used significantly fewer NP-based bundles than expected and British students also used significantly more VP-based bundles and fewer PP-based bundles. If the native expert writing is used as the criterion to judge student writing, the pattern of overuse of VP-based bundles and underuse of NP-based bundles in both L1 and L2 student writing has started to emerge in type distribution, but the extent of this variation is more pronounced in token distribution.

Table 5-13 Chi-square standardised residuals for structural distribution (tokens) in Modular Study 1

$\chi^2=103.0, df=4, p<0.0005$		*NP*	*PP*	*VP*
BAWE-CH	Observed Count	62	206	209
	Expected Count	95.6	185.7	195.7
	R	**-3.4**	1.5	1.0
BAWE-EN	Observed Count	105	205	340
	Expected Count	130.3	253.1	266.6
	R	**-2.2**	**-3.0**	**4.5**
FLOB-J	Observed Count	199	300	200
	Expected Count	140.1	272.2	286.7
	R	**5.0**	1.7	**-5.1**

As can be seen, the cells representing NP-based and VP-based bundles are generally the ones with the higher absolute values of *R* regardless of type or token distribution. By and large, the analysis results in this section suggest that the numbers of NP-based bundles and VP-based bundles can better distinguish groups of writing with different proficiency in the context of academic writing. The evidence from standardised residuals indicates that the distribution pattern of bundle structures in the group with the best writing competency FLOB-J differs from the two student groups, BAWE-EN, and BAWE-CH. The tendency embodied in native professional writing in the main reinforces the quantitative results in Sections 5.3.1 and 5.3.2, which indicate that FLOB-J has more NP-based bundles but fewer VP-based bundles than the two student subcorpora in both type and token distribution. After calculating the standardised residuals, now it is much clearer that in the type distribution, only the categories of NP-based and VP-based bundles in FLOB-J make substantial contributions to the significant difference while in the token distribution, the categories of NP-based and VP-based bundles in two student groups in general also contribute to rejecting the null hypothesis.

What is unexpected is the result of the token distribution (see Table 5-13), in which

one of the cells that make a major contribution to the chi-square statistic is the one representing PP-based bundles in BAWE-EN. Having examined the value of *R*, it is found that BAWE-EN used fewer PP-based bundles than expected. Combined with the results of the type distribution, it can be suggested that British students and Chinese students might manifest a similar range of PP-based bundles, but when the variable of frequency is taken into consideration, British student writing contained significantly fewer occurrences of PP-based bundles while Chinese students used PP-based bundles with higher frequencies. Although the reason of this anomaly still seems unclear here, we shall be able to see what sort of PP-based bundles lead to the underuse in the BAWE-EN in Section 5.6 Keyness Analysis.

Another implication through such analysis is that the proficiency represented in BAWE-EN and in BAWE-CH does not seem to match exactly what was predicted in the beginning, i.e. native peer writing represented in BAWE-EN supposedly being superior to learner writing represented in BAWE-CH and thus showing a pattern which is more similar to native expert writing. By contrast, the results derived from standardised residuals suggest that this group of native student writing shows a more marked difference to FLOB-J than L2 writing in BAWE-CH does. The counter-evidence is particularly striking in the BAWE-EN token distribution when all the three structural categories indicate either underuse (NP-based and PP-based bundles) or overuse (VP-based bundles) whereas in L2 student writing, the only deviation suggested by standardised residuals is the underused NP-based bundles. We shall see whether the assumption that native peer writing is supposed to be of higher quality than L2 writing has fallen apart, at least for the BAWE data being investigated, in further discussion of structural subcategories and functional analysis in the next section.

Searching for the best proficiency indicators is important in the sense that one of the major aims for the present thesis is to describe the distinct linguistic features demonstrated by writers of different proficiency levels. The finding concluded here, heavy reliance on NP-

based bundles and far less reliance on VP-based bundles in professional academic writing, has substantive implications for second language learning and language testing. On the one hand, the two groups of less competent writers, native and non-native students, do not seem to have been aware of the importance of nominal expressions used extensively and frequently by professional writers; instead, they resort to verb phrases or clauses. For teaching and learning of second language writing, therefore, these findings would suggest that more focus should be placed on nominal bundles which function as markers for academic writing. In the overall discussion chapter in this thesis, we shall see how the textbooks of academic writing have failed to achieve this. On the other hand, such distinctive difference in the use of NP-based and VP-based bundles, to my knowledge, does not seem to have been reported in the studies of language development or language testing.[24] This is therefore an area which is very much worthy of further research.

5.3.4 Lexical Bundles in Structural Categories

Remember that all the bundles have been assigned into a specific subcategory under three major structural categories: NP-based, PP-based, and VP-based (cf. Section 4.2). Over 40% of the subcategories, unfortunately, contain fewer than five instances in the three corpora because of the fine-grained structural categorisation system, and hence the chi-square test cannot be applied to the subcategory distribution. Some patterns of the usage of lexical bundles, nevertheless, can still emerge under close scrutiny of structural categories.

---

[24] Comparing 'advanced' learner writing and native student writing (i.e. ICLE and LOCNESS), Granger and Rayson (1998) did point out that learners overused determiners, pronouns and adverbs significantly and underused conjunctions, prepositions and nouns significantly ($p<0.01$). Yet with a three-way comparison (non-native student writing, native peer student writing, and native expert writing) and a phraseological approach, the results presented in this thesis are thus very different from Granger and Rayson's study.

5.3.4.1 NP-based and PP-based Bundles

NP-based and PP-based bundles are grouped together for investigation. One reason for this decision is that neither one of them contains a verb component. Another is that a preposition often precedes a NP-bundle while PP-based bundles always have a noun or noun phrase following the preposition. To begin with, NP-based bundles are made up of two subcategories: nominal phrase fragments with *of* (NP+*of*) and any other nominal phrase fragments without *of* (NPf); likewise, PP-based bundles comprise prepositional phrase fragments with *of* (PP+*of*) and any other prepositional phrase fragments without *of* (PPf). In addition to the comparably low proportion of NP-based bundles when compared with FLOB-J, Chinese student writing represented in BAWE-CH also distinguishes itself from the two groups of native writing in the subcategory of NPf because there is no such defined word combination in BAWE-CH which falls in this subcategory. On the contrary, the NPf bundles appearing in the professional FLOB-J writing are mostly used by the British students in BAWE-EN, although there is some slight variation (see Table 5-14). NPf combinations found in this investigation are all part of a relative clause such as *the extent to which*, *the fact that this*, or *the way(s) in which*. It is evident that Chinese L2 students did not use this type of relative clause as frequently as native speakers did.

Table 5-14 NPf bundles shared by FLOB-J and BAWE-EN

FLOB-J	Freq	BAWE-EN	Freq
the degree to which	5	***the extent to which***	8
the extent to which*	6		
the fact that this	4	*the fact that the*	8
		the fact that they	4
the way in which	14	***the way in which***	7
the ways in which	4		
type	5	**type**	4
token	33	**token**	27

*Bundles appearing in both native corpora are indicated in bold print.

Secondly, a great number of NP+of and PP+of bundles can be grouped into two productive frames: '*the* + Noun + *of the/a*' and '*in the* + Noun + *of*'. The professional writing in FLOB-J manifests a relatively wide range of nouns which fit in these two fixed frames (Table 5-15 and Table 5-16). In this regard, the patterns emerging here lend support to the finding reported by Biber et al. (2003), who described the same two 'fixed frames' (or termed as 'phrase-frame' by Stubbs, 2007a) used for 43 and 17 different types of lexical bundles respectively in their academic prose of five million words as highly productive. In contrast, neither the English L1 students nor the Chinese L1 students seem to have recognised the importance of these nominal or prepositional expressions in academic writing.

Table 5-15 The frame '*the* + Noun + *of the/a*' used in Modular Study 1

the* + Noun + *of the/a		**Total**	
		type	**token**
BAWE-CH	**development** (4), **end** (4), importance (4), **nature** (5), **rest** (8), **role** (4), **size** (4), top (4)	8	37
BAW-EN	**development** (11), **end** (10), length (6), **nature** (7), quality (4), **rest** (12), **size** (4), **structure** (6), use (9)	9	69
FLOB-J	**end*** (10), creation (4), existence (4), history (7), impact (4), magnitude (4), results (4), **nature** (17), **rest** (11), **role** (5), rules (5), **size** (7), status (4), strength (5), **structure** (4), value (5)	16	100

* Lexical bundles appearing in two or three corpora are indicated in bold print. In addition, the frequency is indicated in the brackets. The same practice is used for Table 5-16.

Table 5-16 The frame '*in the* + Noun + *of*' used in Modular Study 1

in the* + Noun + *of		**Total**	
		type	**type**
BAWE-CH	**case** (10), **context** (5), **form** (4)	3	19
BAWE-EN	**absence** (4), **case** (23), **form** (8)	3	35
FLOB-J	**absence** (7), **case** (19), **context** (19), course (5), face (4), **form** (8), hands (5), light (6), number (6), presence (8)	10	87

Not only did the student writers demonstrate a relatively limited range of interchangeable nouns that go with the two fixed frames, more specifically, the British student writing in BAWE-EN could also be characterised with an unusual repetitive appearance of the noun *use* incorporated in NP+*of* and PP+*of* bundles. There are four such bundles: *the use of the*, *and the use of*, *through the use of*, and *for the use of*, none of which are shared by the other two corpora in the target bundles. In fact, even more VP-based bundles with the verb form '*used*' were found in BAWE-EN: *be used in the*, *can be used to*, *could be used to*, *can be used for*, *also be used to*, *and can be used*, *can also be used*, and *will be used to*. Only one of the above, *can be used to*, also appeared in FLOB-J and in BAWE-CH, and its frequencies in both corpora are lower than in BAWE-EN. All the lexical bundles with *use/used* as the core element in BAWE-EN add up to 11.5% of the total bundle types in the British student writing. This extraordinary persistence with regard to the use of *use/used* in BAWE-EN is quite perplexing. Even after the concordance lines of such instances have been checked, there still seems no clear explanation for this unusual preference as these expressions with *use/used* are widely distributed across disciplines and used by different native student contributors. In the following examples with the expression *the use of the* retrieved from British student writing, however, the usage does appear either awkward, as in the first example, or redundant, as in the second example.

- It is, however, very important that scoping is carried out with ***the use of the*** legislative requirements and good practice in order to limit both short and long-term damage to the environment. (BAWE-EN)

- This explanation does indeed appear to be simpler, and if ***the use of the*** principal of Occam's Razor is applied, the postulation of God should therefore be the idea followed. (BAWE-EN)

On the other hand, one of the possible reasons why there are relatively fewer expressions with *use/used* in published academic writing might be that the professional academic writers, with a wider range of lexical complexity and variation, know how to distinguish the nuances of meanings and choose another word for *use/used* to express themselves more precisely since there are more than ten entries of explanation under the noun/verb *use*.[25] It is also possible that the professional academic writers are able to make use of other expressions to avoid repetition in their writing, so they would find alternatives to replace *use* or *used* when it has just appeared in the previous text, as in the following example.

- For serial transmission or storage, start and stop bits ***are used to*** delimit a data word, and so either parity type ***can be employed.*** (FLOB-J)

The assumption above could be evidenced to some extent by the standardised type/token ratios (STTR) as having been discussed in Section 5.2.1 Type/token Ratio. Both sets of student writing have lower STTRs than the professional writing in FLOB, although BAWE-EN is even slightly lower than BAWE-CH (39.16% vs. 39.3%, see Table 5-7). In other words, British students represented in BAWE-EN, unexpectedly, demonstrated the smallest range of lexis while native scholars represented in FLOB-J exhibited the greatest range of vocabulary use. Still, the fact that Chinese students display a slightly wider range of

---

[25] See the American Heritage Dictionary (2003), cited from <http://www.thefreedictionary.com> (visited on 12 May, 2009).

vocabulary than British students might not be sufficient alone to respond to the question of whether Chinese student writing is closer to native expert writing than British student writing is as revealed by structural distribution of bundles addressed in Section 5.3.3. The frequency of the orthographic forms *use/used* in BAWE-CH, as a matter of fact, is higher to a small degree than that in BAWE-EN (see Table 5-17), and the keyword analysis confirms that Chinese students also overused *use/used* when compared to native academic writers. The results of having checked the concordance lines of *use/used* in Chinese students' writing suggest that Chinese student writers tend to incorporate *use/used* in shorter word sequences such as *the use of*, *use of direct*, *language use*, *that the use*, *be used to*, *can be used*, *used in the*, *are used to*, *used as a*, etc., but not in the recurrent four-word sequences, at least the frequency of which did not reach the threshold set in the current study. Incidentally, the twelve bundles with *use/used* in BAWE-EN can be foregrounded to explain why the number of total lexical bundles in BAWE-EN (104 bundle types in total) is a lot higher than that in BAWE-CH (80 bundle types in total), as the two groups of student writing have displayed similarity in many aspects. The discussion concerning *use/used* or STTR in this section might look a digression away from the theme of lexical bundles. Yet it provides some valuable information about the linguistic features of the three corpora and also perhaps some answers to the enquiries, e.g. why the number of bundles retrieved from BAWE-EN is much higher than that in BAWE-CH as they were expected to be much closer in this regard.

Table 5-17 Frequency of the orthographic forms *use/used* in Modular Study 1

Frequency	*BAWE-CH*	*BAWE-EN*	*FLOB-J*
use/used	435	427	253

5.3.4.2 VP-based Bundles

With reference to the subcategories in VP-based bundles, the student writers in BAWE-EN and BAWE-CH used particularly more 'Pronoun/Noun + BE/Verb phrases' (see Table 5-18) and '*to*-clause fragments' (see Table 5-19). In these two subcategories, there is more similarity between BAWE-EN and BAWE-CH.

Table 5-18 The subcategory of 'Pronoun/Noun + BE/Verb phrases' in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
essay is going to	4	*an example of this*	8	*there is evidence that*	5
that is to say	6	***that is to say****	4		
this is due to	4	*that need to be*	4		
this means that the	4	*there is no evidence*	4		
we can see that	7	*there would be no*	5		
		this is due to+(the)	5		
		this may be due to	4		
		this means that the	4		
type	5	**type**	8	**type**	1
token	25	**token**	38	**token**	5

* The bundles appearing in two or three corpora are indicated in bold. The same practice is used for the rest of this thesis.

Table 5-19 The subcategory of '*to*-clause fragments' in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
in order to achieve	8	***in order to make***	8	***to be able to***	5
in order to avoid	7	*in order to minimise*	4		
in order to be	5	***to be able to***	8		
in order to maintain	4	*to be added to*	4		
in order to make	4	*to cope with the*	4		
in order to understand	4	*to enable them to*	4		
to be able to	4	*to take into account*	4		
to ensure that the	4				
type	8	**type**	7	**type**	1
token	40	**token**	36	**token**	5

Here we shall start with 'Pronoun/Noun + BE/Verb phrases'. Take *this is due to* for example. Both groups of student writers used it to convey the inferential sense, as the

following two examples illustrate. The British student in the first example used *this is because* and *this is due to*, two very similar expressions, in two consecutive sentences, which might sound awkward.

- This is because there has been a dramatic decline in the numbers of native UK species, including the red squirrel, pine marten, water vole and common dormouse. ***This is due to*** a change in agricultural systems affecting habitats and feeding site, competition with other native and non-native species, and previous hunting pressures over the previous two centuries. (BAWE-EN)

- A marked decrease started in the late 1940s, dropping to its neap, 1.4 in 1960. ***This is due to*** the introduction of DDT into the agricultural use at that time. (BAWE-CH)

In the subcategory of '*to*-clause fragments', student writers, particularly Chinese students, appear to favour the frame '*in order to* + Verb', particularly using six different verbs that fit in the slot: *achieve*, *avoid*, *be*, *maintain*, *make*, and *understand* while British student writing had two such bundle types: *in order to make* and *in order to minimise*. It would be wrong, however, to accordingly assume that the professional writers did not use the expression *in order to* in their writing. Actually there are 30 occurrences of '*in order to*' in FLOB-J, but none of the word combinations meet the frequency threshold. In comparison, both student groups used this expression far more than just 30 times: 97 occurrences in BAWE-EN and 125 in BAWE-CH (see Table 5-20). This, again, could be the result of professional writers' mastery of varying such expressions by virtue of their wider range of lexis. Furthermore, a finding like this, once more, illustrates the power of an approach based on *multi-word sequences* that a traditional analysis based on *words* cannot achieve.

Table 5-20 Frequency of *in order to* in Modular Study 1

Frequency	*BAWE-CH*	*BAWE-EN*	*FLOB-J*
in order to	125	97	30

Although there are a substantial number of VP-based bundles in BAWE-CH, 50% within the corpus by bundle types and 43.8% within the corpus by bundle tokens, Chinese students did not appear to use the 'Passive verb + prepositional phrases' (PassPP) forms as frequently as native speakers did. As can be seen in Table 5-21, there are seven different passive-verb bundles in FLOB-J and eleven in BAWE-EN, both of which take around 20% in the category of VP-based bundles (types) respectively. By contrast, the four different passive-verb bundles in BAWE-CH constitute merely 10% of the total VP-based bundles (types), not to mention that none of the four passive bundles was shared by either of the native groups of writers. It was also found that British students seem to favour the expression *take into account*, which has three variants in the lexical bundles: *be taken into account*, *to take into account*, and *taking into account the*. Only *be taken into account* is found in FLOB-J while none of these bundles occur in BAWE-CH.

Table 5-21 The subcategory of 'Passive verb +prepositional phrases' in Modular Study 1

BAWE-CH	Freq	**BAWE-EN**	Freq	**FLOB-J**	Freq
can be divided into	4	*+be seen as a*	5	*are shown in fig*	6
can be explained by	7	*be included in the*	4	*be found in the*	5
can be regarded as	4	***be taken into account***	5	*be seen in the*	4
is illustrated in figure	4	*be used in the*	5	***be taken into account***	5
		can be applied to	7	***can be found in***	6
		can be found in	6	*is concerned with the*	4
		can be seen as+	5	*was followed by a*	4
		can be seen in	4		
		can be used for	5		
		could be seen as	5		
		should be placed on	4		
type	4	**type**	11	**type**	7
token	19	**token**	55	**token**	34

The category of PassPP, however, does not include all the bundles with any passive verbs as some such bundles were categorised into other subcategories according to the rules of structural categorisation addressed in Section 4.2.3, e.g. *could be used to* in 'VP + *to*-clause' and *it is believed that* in '*it* pattern'. However, a careful examination still suggests the general tendency that Chinese students used far fewer passive forms although there does not seem to be any noticeable pattern among the three corpora in the passive-verb bundles within the subcategories of 'VP + *to*-clause' (see Table 5-22) and '*it* pattern' (see Table 5-23).

Table 5-22 Passive-verb bundles in the subcategory of 'VP + *to*-clause' in Modular Study 1

BAWE-CH	Freq	**BAWE-EN**	Freq	**FLOB-J**	Freq
can be used to	10	*can also be used+(to)*	5	***can be used to***	6
is considered to be	4	***can be used to***	17		
		could be used to	6		
		will be used to	4		
type	2	**type**	4	**type**	1
token	14	**token**	32	**token**	6

Table 5-23 Passive-verb bundles in the subcategory of '*it* pattern' in Modular Study 1

BAWE-CH	Freq	**BAWE-EN**	Freq	**FLOB-J**	Freq
it can be seen	4	*it can be argued+(that)*	5	***it can be seen that***	5
it has been suggested that	6	***it can be seen + (that)***	12	***it has been suggested***	4
it is believed that	5	*it could be argued that+ (the)*	14		
		it is estimated that	4		
		it should be noted	4		
type	3	**type**	5	**type**	2
token	15	**token**	39	**token**	9

Overall, the two groups of student writing appear to be closely similar in terms of bundle structures, sharing a number of identical patterns such as having very few bundles in the two fixed frames '*the* + Noun + *of the/a*' (e.g. *the nature of the*) and '*in the* + Noun + *of*

(e.g. *in the context of*) or favouring bundles with the '*in order to* + VP' structure (e.g. *in order to make*). On the other hand, learner writing also distinguishes itself from native writing not only by the lack of NPf bundles (e.g. *the extent to which*) but also by containing fewer passive-verb bundles (e.g. *can be found in*). In the next section, we shall see how the three groups of writing differ from or resemble each other in terms of functional analysis.

5.4 Functional Analysis

In functional analysis, each bundle has been assigned to a corresponding functional category according to the discourse function it plays (referential expressions, stance bundles, or discourse organisers, cf. Section 4.3.3). Type and token distributions are likewise distinguished. Unlike structural analysis, no category of 'Others' needed to be created in the functional categorisation to cater for any miscellaneous bundles; that is to say, each lexical bundle is included in the quantitative analysis because each of them has been assigned to an appropriate category.

5.4.1 Type Distribution

As can be seen from Table 5-24 and Figure 5-3, referential expressions are the majority in FLOB-J (60.2%) whereas they are a lot less frequent in both BAWE-EN (36.5%) and BAWE-CH (41.3%). On the other hand, discourse organisers rank as the largest category in both BAWE-EN and BAWE-CH with very similar proportions of 39.4% and 42.5% respectively while discourse organisers in FLOB-J take only about half of that (21.3%). As to stance bundles, BAWE-EN has the highest percentage of use, 24%, but BAWE-CH and FLOB-J both rely on stance bundles to a lesser degree (16.3% and 18.5% respectively).

Table 5-24 Functional distribution in Modular Study 1 (types)

χ^2=16.4, df=4 p=0.003			Function			Total
			Referential expressions	Stance bundles	Discourse organisers	
Corpus	BAWE-CH	Count	33	13	34	80
		% within Corpus	41.3%	16.3%	42.5%	100.0%
	BAWE-EN	Count	38	25	41	104
		% within Corpus	36.5%	24.0%	39.4%	100.0%
	FLOB-J	Count	65	20	23	108
		% within Corpus	60.2%	18.5%	21.3%	100.0%

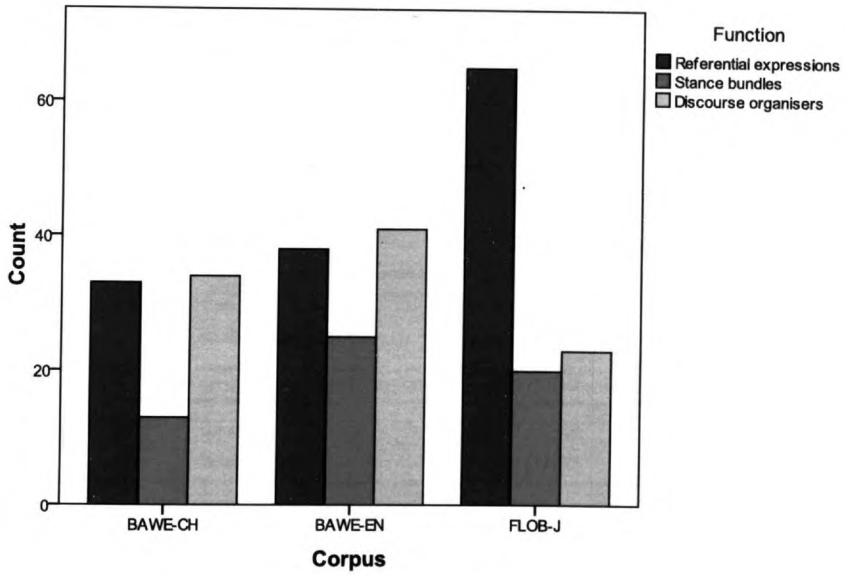


Figure 5-3 Functional distribution in Modular Study 1 (types)

The chi-square test indicates that there is significant difference in terms of functional distribution of bundle types between BAWE-CH, BAWE-EN, and FLOB-J (χ^2=16.4, df=4 p=0.003) at the significance level 0.05. The standardised residuals will be calculated later to figure out which functional categories make a major contribution to rejecting the null hypothesis.

5.4.2 Token Distribution

The token distribution of functions among the three corpora is virtually the same as the type distribution. As can be seen in Table 5-25 and Figure 5-4, the proportion of referential expressions remains the most marked difference among the three groups of writing, as referential expressions take almost two thirds of the bundles in FLOB-J while BAWE-CH and BAWE-EN both contain fewer referential expressions (38.7% and 36.9% respectively). Instead, the two student groups of writing rely more heavily on discourse organisers. Chinese student writing in BAWE-CH particularly demonstrates an unusually high percentage of discourse organisers as high as 48.1%.

Table 5-25 Functional distribution in Modular Study 1 (tokens)

χ^2=148.5, df=4, p<0.0005			Function			Total
			Referential expressions	Stance bundles	Discourse organisers	
Corpus	BAWE-CH	Count	196	67	244	507
		% within Corpus	38.7%	13.2%	48.1%	100.0%
	BAWE-EN	Count	246	161	260	667
		% within Corpus	36.9%	24.1%	39.0%	100.0%
	FLOB-J	Count	437	125	142	704
		% within Corpus	62.1%	17.8%	20.2%	100.0%

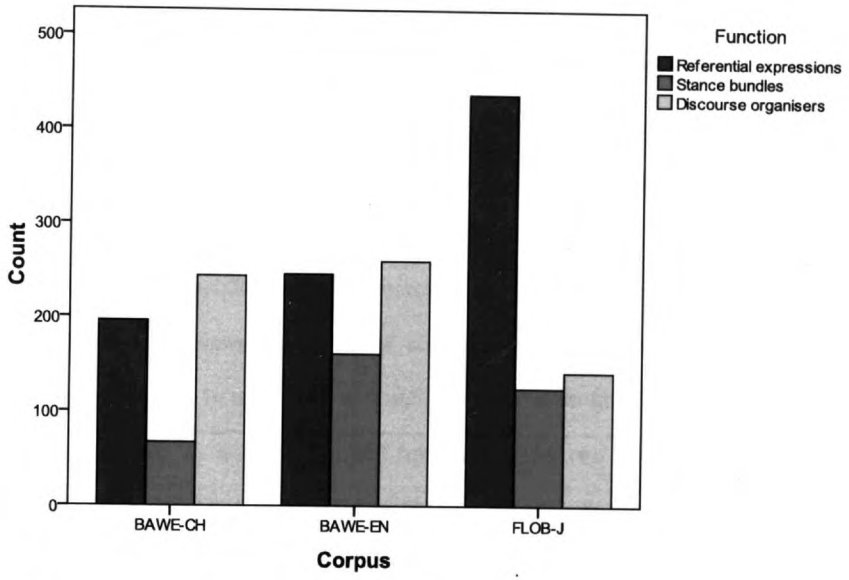


Figure 5-4 Functional distribution in Modular Study 1 (tokens)

The chi-square test indicates that there is significant difference in terms of functional distribution of bundle tokens between BAWE-CH, BAWE-EN, and FLOB-J (χ^2=148.5, df=4, p<0.0005).

Interestingly enough, the results of significance tests based on functional distribution mirrors exactly that of the structural distribution in terms of type and token comparison. No matter whether it involves structural or functional analysis, the value of chi-square statistic χ^2 is much greater in token distribution than in type distribution, which suggests that the extent of variation turns to be much more pronounced when the variable of frequency is taken into account. This is not surprising as chi-square statistics are also dependent on sample size. Moreover, there is a high correlation between structural and functional categorisation, which may very much contribute to the corresponding relationship in bundle distribution between the three corpora. This is further discussed in Section 5.5.

5.4.3 The Chi-square Test & Standardised Residuals

Following the rationale behind the chi-square test discussed in Section 5.3.3, the chi-square test was again executed with the three corpora. This time the focus would be cast on which cells in the three (corpus) by three (functional category) table contribute to the significant difference among the three groups of writing so that they can be interpreted as better candidates for distinguishing writing competency.

Table 5-26 shows the results of standardised residuals *R*, which can be used to compare which cells in the matrix of functional distribution in terms of bundle types make more contribution to rejecting the null hypothesis. The two cells representing discourse organisers and referential expressions in FLOB-J have the absolute value of *R* higher than 1.96 (2.2 and 2.1 respectively), thereby being the two cells that bear the most responsibility for the statistical difference found. An overall observation indicates that native expert writing contains more referential expressions yet fewer discourse organisers than expected whereas the two student groups of writing show the opposite pattern, i.e. with fewer referential expressions yet more discourse organisers than expected.

Table 5-26 Chi-square standardised residuals for functional distribution (types) in Modular Study 1

$\chi^2=16.4$, $df=4$, $p=0.003$		***Referential expressions***	***Stance bundles***	***Discourse organisers***
BAWE-CH	Observed Count	33	13	34
	Expected Count	37.3	15.9	26.8
	R	-0.7	-0.7	1.4
BAWE-EN	Observed Count	38	25	41
	Expected Count	48.4	20.7	34.9
	R	-1.5	0.9	1.0
FLOB-J	Observed Count	65	20	23
	Expected Count	50.3	21.5	36.2
	R	**2.1**	-0.3	**-2.2**

The same two cells, referential expressions and discourse organisers in FLOB-J, remain the top two in the functional distribution in terms of bundle tokens, the value of *R* reaching as high as 5.9 and as low as -6.4 (see Table 5-27). The cell representing discourse organisers in BAWE-CH has the third highest absolute value of *R*, 5.3, which confirms the descriptive statistics that L2 students used notably more discourse organisers than native academicians. It is also interesting to note that all the three functional categories in both student corpora are indicative of major contributors to the significant difference. The values of *R* suggest that two student groups of writing generally displayed overuse and underuse in the same pattern: underuse of referential expressions yet overuse of discourse organisers. Only the category of stance bundles is an exception. BAWE-CH contains fewer stance bundles than expected whereas BAWE-EN has more stance bundles than expected. Combined with the results of type distribution (cf. Table 6-26), it can be suggested that British students and Chinese students used a similar range of stance bundles, but when the variable of frequency is taken into consideration, British student writing tended to employ stance bundles with significantly higher frequencies whereas Chinese students tended to use them far less frequently. In the next section, we will see what sort of expressions contributed to the underuse of stance bundles in Chinese student writing.

Table 5-27 Chi-square standardised residuals for functional distribution (tokens) in Modular Study 1

$\chi^2=148.5$, $df=4$, $p<0.0005$		***Referential expressions***	***Stance bundles***	***Discourse organisers***
BAWE-CH	Observed Count	196	67	244
	Expected Count	237.3	95.3	174.4
	R	**-2.7**	**-2.9**	**5.3**
BAWE-EN	Observed Count	246	161	260
	Expected Count	312.2	125.4	229.4
	R	**-3.7**	**3.2**	**2.0**
FLOB-J	Observed Count	437	125	142
	Expected Count	329.5	132.3	242.2
	R	**5.9**	-0.6	**-6.4**

The chi-square residuals in Table 5-26 and Table 5-27 manifest a general relationship of functional distribution of lexical bundles in the three corpora. The results in this section as a whole suggest that the numbers of referential expressions and discourse organisers can better distinguish groups of writing with different proficiencies in the context of academic writing. In both type and token distribution, they indicate the group of writing with the best writing competency, FLOB-J, by means of the highest absolute values of *R* in these two functional categories. The tendency exhibiting in native professional writing in the main reinforces the quantitative results in Sections 5.4.1 and 5.4.2, which concludes that FLOB-J overall has more referential expressions but fewer discourse organisers, with an intermediate percentage of stance bundles when compared with the two student subcorpora. One similar pattern shared by functional and structural analysis (cf. Section 5.3.3) is that token distribution appears to always demonstrate a more pronounced extent of variation in comparison with type distribution, which lends support to the practice of distinguishing bundle types and tokens in quantitative analysis.

Another important finding concerns the proficiency represented in BAWE-EN and in BAWE-CH. Remember in the structural analysis (Section 5.3), the assumption that native peer writing represented in BAWE-EN is supposed to be superior to learner writing represented in BAWE-CH and therefore closer to FLOB-J was challenged in the light of the results of chi-square residuals. The functional analysis here somewhat confirms the rebuttal initialled by the structural analysis on the ground that the values of standardised residuals BAWE-EN and BAWE-CH generally display a similar distribution pattern except for the category of stance bundles. Combined with the structural analysis, it then appears that the usage of lexical bundles is similar in native peer writing and learner writing irrespective of bundle structures or functions, at least for the BAWE data being investigated with the current quantitative approach. The difference is only prominent when frequency is taken into account.

In the following section of functional categories, we shall see whether this high degree of similarity lingers between these two groups of student writing.

5.4.4 Lexical Bundles in Functional Categories

In this section, the use of lexical bundles in functional subcategories under each major category (cf. Section 4.3.3) will be discussed and exemplified with some extended concordance lines extracted from the corpora with a view to further examining the use of lexical bundles in terms of discourse functions. In addition, the graphs which show the breakdown of each major functional category will be presented to provide a better overview, yet only the numbers of bundle types are presented in the graphs as this thesis does not intended to bombard the readers with dozens of graphs including both bundle types and tokens. Meanwhile, the distribution patterns in types and token are generally similar, although we already know that the latter often displays a more pronounced pattern of variation.

5.4.4.1 Referential Expressions

At a first glance at the breakdown of referential expressions (Figure 5.5), it can be seen that the professional writers in FLOB-J used the largest range of different types of bundles in each of the functional subcategories of referential expressions, but it seems that the three groups of writing do not differ much in the distribution of referential bundles used for specifying attributes ('framing bundles', e.g. *in the context of*), expressing quantity, amount or extent ('quantifying bundles', e.g. *a great number of*), and time/place/text deixis ('deictic bundles', e.g. *at the same time*).[26] We shall see whether the groups of writers used referential bundles in a distinctive fashion by a further examination of individual bundles and their usage below.

---

[26] A chi-square test indicates that there is no significant difference among the three corpora in terms of the distribution of referential subcategories (χ^2=1.9, df=4, p=0.749). This is the only functional category which can be analysed with a chi-square test because every referential subcategory contains no fewer than five instances.

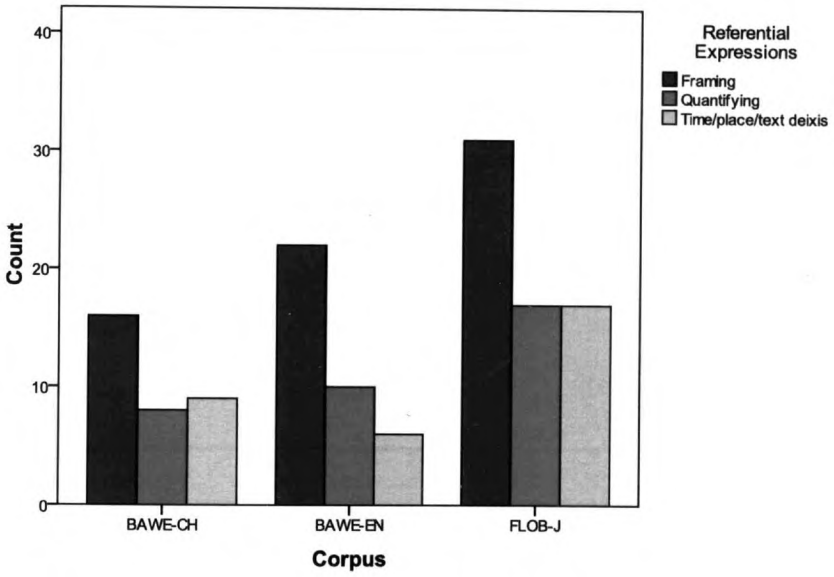


Figure 5-5 Breakdown of referential expression in Modular Study 1 (types)

With regard to the referential framing bundles (Table 5-28), many of them have been discussed earlier with the two frames '*the* + Noun + *of the/a*' and '*in the* + Noun + *of*' in Section 5.3.4.1. If we focus on the lexical bundles that are shared by native expert writing and either of the student corpora, the British student writers in BAWE-EN are found to share ten bundle types with the native scholars in FLOB-J while there are nine such shared bundle types between BAWE-CH and FLOB-J. This proportion of overlapping is actually fairly high, especially in BAWE-CH, which has 56.3% referential framing bundles shared with FLOB-J while BAWE-EN has 45.5% framing bundles shared with FLOB-J. In this regard, both student groups seem to have started grasping certain sense of framing expressions in the desired style of academic writing, although the range is not as wide as native expert writing.

Table 5-28 Referential framing bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
as a part of	5	***(for)+the development of the***	13	*a function of the*	5
as part of a	5	*an integral part of*	4	*a function of time*	6
at the expense of	6	*and the use of*	6	*as a function of*	15
for the development of	4	*as a way of*	4	*by a variety of*	4
in terms of the	5	*by the presence of*	4	***in relation to the***	4
in the case of	10	*for the use of+*	4	*in terms of a*	4
in the context of	5	***in relation to the***	5	***in terms of the***	14
in the form of	4	*in terms of its*	5	***in the absence of***	7
on the basis of	5	***in terms of the***	13	***in the case of***	19
the development of the	4	***in the absence of***	4	***in the context f+(the/a)***	19
the importance of the	4	***in the case of***	23	*in the face of*	4
the nature of the	5	***in the form of***	8	***in the form of***	8
the role of the	4	*in the same way*	9	*in the hands of*	5
to the development of	6	***the nature of the***	7	*in the presence of*	8
with respect to the	4	*the quality of the*	4	***on the basis of***	14
with the introduction of	4	***the structure of the***	6	*on the part of*	6
		+the use of the	9	*the creation of a*	4
		the way in which	7	*the existence of a*	4
		through the use of	5	*the history of the*	7
		with respect to the	6	*the impact of the*	4
		with the addition of	4	***the nature of the***	17
		with the development of	5	*the point of view*	4
				the role of the	5
				the rules of the	5
				the status of the	4
				the strength of the	5
				the structure of the	4
				the value of the	5
				the way in which+(the)	14
				the ways in which	4
				with respect to the	6
type	16	**type**	22	**type**	31
token	80	**token**	155	**token**	234

A moderate proportion of overlap is maintained in the subcategories of quantifying (Table 5-29) and deictic bundles (Table 5-30), yet both groups of student writing, again, always exhibit a narrower range of bundles.

Table 5-29 Referential quantifying bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
a high level of	4	*a great deal of*	4	***a high level of***	4
a large number of	4	***as part of the***	4	***a large number of***	4
a number of factors	6	*for each of the*	5	***a wide range of***	9
a wide range of	4	*of some of the*	4	*for each of the*	5
as part of the	4	***of the number of***	5	*has a number of*	4
of the number of	4	***the extent to which***	6	*in a number of*	5
the rest of the	8	*the length of the*	6	*in so far as*	6
the size of the	4	***the size of the***	4	*in the number of*	6
		the rest of the	12	***of a number of***	6
		to a certain extent	4	*on a number of*	4
				per cent of the	9
				the degree to which	5
				the extent to which	8
				the size of the	7
				the magnitude of the	4
				the rest of the	11
				to a large extent	4
type	8	**type**	10	**type**	17
token	38	**token**	50	**token**	97

One type of quantifying bundles is noteworthy, i.e. the extent/degree modifiers, which are present in native writing but not in learner writing (see Table 5-29). There are four such bundles in FLOB-J: *in so far as*, *the degree to which*, *the extent to which*, and *to a large extent* while there are two in BAWE-EN: *the extent to which* and *to a certain extent*. It appears that learners did not use these modifiers much whereas native speakers tended to use them to modify the extent or degree of their proposition as the following examples show.

- No matter what the nature of the being, the principle of equality requires that its suffering be counted equally with the like suffering - ***in so far as*** rough comparisons can be made - of any other being. (FLOB-J)
- ***The degree to which*** these effects are found in normal reading has recently been examined by McConkie, Kerr, Reddix and Zola (1988). (FLOB-J)

- Thus even though when one entertains in commercial setting aspects of intimacy can work well ***to a certain extent.*** (BAWE-EN)

Table 5-30 Referential deictic bundles in Modular Study 1

BAWE-CH	Freq	**BAWE-EN**	Freq	**FLOB-J**	Freq
(and)+at the same time	24	***(at)+the end of the***	13	***(at)+the end of the***	13
all over the world	6	*at the heart of*	4	*are shown in fig*	6
at the beginning of	6	***at the same time***	5	*as shown in fig*	6
at the end of+(the)	8	*be used in the*	5	*as we have seen*	8
in the end of+(this)	7	***can be found in***	6	*as we shall see*	7
in the long run	13	*can be seen in*	4	*at each end of*	4
in the recent years	6			***at the beginning of***	4
is illustrated in figure	4			*at the time of*	5
the top of the	4			***at the same time***	10
				be found in the	5
				be seen in the	4
				can be found in	6
				end of the spectrum	4
				in the course of	5
				the right hand side	4
				the second half of	4
				the turn of the century	7
type	9	**type**	6	**type**	17
token	78	**token**	37	**token**	102

On the other hand, Chinese student writers seem to use certain referential deictic expressions that were not used as frequently as in published academic writing. As can be seen in Table 5-30, some of the frequent deictic expressions used in Chinese student writing, *in the long run*, *in the recent years*, and *all over the world*, as exemplified in the following examples, do not appear in the target bundles used by the professional writers nor the British peer students.

- Almost all economists today agree that monetary policy influences unemployment, at least temporarily, and determines inflation, at least ***in the long run.*** (BAWE-CH)
- Education may require individuals to forgo current earnings as opportunity cost in the short run; but ***in the long run,*** a skilful

labour force can increase productivity and efficiency, which can lead to an increase in output. (BAWE-CH)

- They are more or less equivalent way of paying out retained earning, while stock repurchases indeed have become an important source of payout ***in the recent years***. (BAWE-CH)
- Cafeteria Style Fringe Benefits (Flexible benefits): This strategy is now very popular ***all over the world***, for it maximizes the value of limited monetary amount of fringe benefits and gives the employees some controls over their own rewards. (BAWE-CH)
- In this way so open-source can be proposed as a bridge for the technological, educational and cultural gaps between developing and developed countries. It allows collaboration and cooperation with a wide spectrum of experts in high-tech fields ***all over the world***. (BAWE-CH)

The first word combination, *in the long run*, is an idiomatic expression, occurring 13 times in BAWE-CH while only once in FLOB-J and twice in BAWE-EN. This idiom *in the long run* is actually more characteristic of non-academic text than of academic prose, and is also quite frequent in speech, as indicated by the British National Corpus (BNC)[27], albeit not being identified as an informal expression in dictionaries (e.g. Macmillan Dictionary for Advanced Learners, Rundell, 2007). The second bundle, *in the recent years*, was generally expressed as *in (more) recent years* and *recently* by native writers in FLOB-J and BAWE-EN. Interestingly, we found 2,344 instances of *in recent years* and only 2 instances of *in the recent years* in BNC. A possible explanation is that *in the recent years* is distinctive to Chinese students due to the redundancy of the article *the*. The third expression, *all over the world*,

[27] In the BNC, for academic writing, the frequency per million words of *in the long run* is 6.72. This figure is 8.27 for non-academic prose and 4.23 for speech.

might reflect the tendency of learners to be categorical and over-generalising as this expression seems to be favoured by learners of various proficiency levels, as we shall see in the next chapter. It might also be concerned with a similar learner idiosyncrasy, i.e. the absence of extent/degree modifiers, which has been discussed in this section and will be further revealed in the discussion of stance bundles.

5.4.4.2 Stance Bundles

Two broad subcategories are distinguished in stance bundles: those used to express the writer's epistemic evaluation of a proposition in terms of its certainty, probability or possibility (epistemic) and those used to convey the writer's attitude towards the proposition in terms of desire, obligation/directive, intention, and ability (attitudinal/modality). As can be seen from Figure 5-6, the professional academic writers in FLOB-J used the most different types of epistemic expressions but the least different types of attitudinal/modality bundles while the Chinese student writing in BAWE-CH exhibited a relatively small range of different epistemic bundles. The native student writing in BAWE-EN contains the most attitudinal/modality bundles and also quite numerous epistemic bundles.

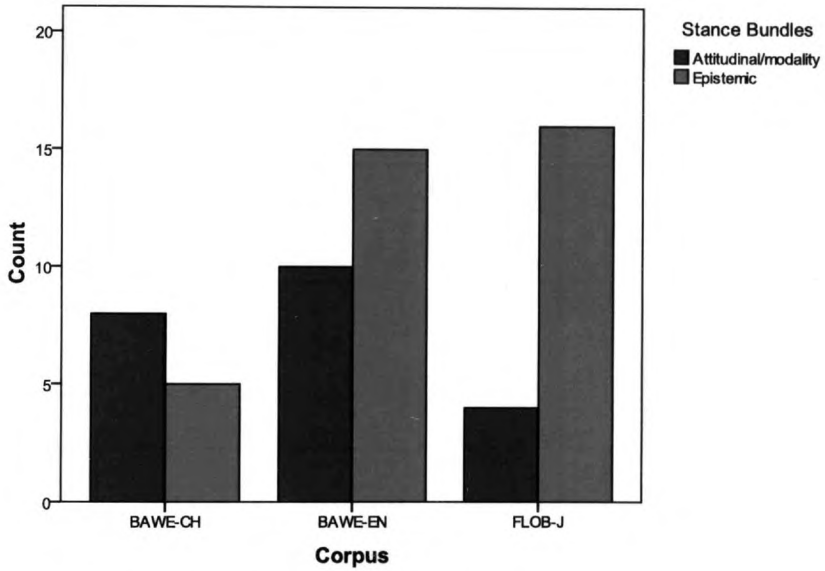


Figure 5-6 Breakdown of stance bundles in Modular Study 1 (types)

If we turn to the stacked bar chart presented in Figure 5-7, which indicates the proportion of epistemic use and attitudinal/modality use, there then appears to be a linear relationship with the three groups of writers in terms of proportional use of sub-functions in stance bundles. Combined with the visual representation shown in Figure 5-6, the supposedly least competent writers, represented by the student writers in BAWE-CH, relied on a very limited range of epistemic bundles at their disposal yet a wider range of attitudinal/modality bundles, despite the frequency of both being comparably low. On the contrary, the most proficient writers in FLOB-J manifested an opposite pattern.

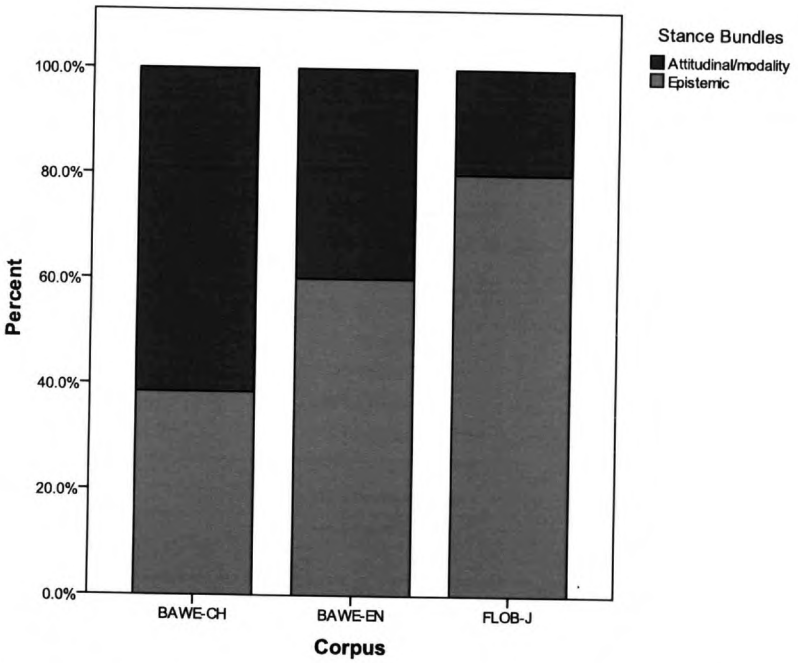


Figure 5-7 Proportional use of breakdown of stances bundles in Modular Study 1 (types)

A further investigation of epistemic markers used by the native writers showed that both groups of native writers are fairly capable of taking advantage of comprehensive measures to control the degree of certainty in their statements (16 and 15 different types of epistemic bundles in FLOB-J and BAWE-EN respectively). In contrast, Chinese student writing contained only five different types of epistemic bundles for this purpose (see Table 5-31).

Table 5-31 Stance epistemic bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
are more likely to	5	***are likely to be***	5	***are likely to be***	4
it has been suggested that	6	***are more likely to***	6	***are more likely to***	7
it is believed that	5	*is by no means*	4	*by the fact that*	9
is considered to be	4	*is the fact that*	6	*is likely to be*	7
to the fact that	4	*it can be argued+(that)*	5	***it has been suggested***	4
		it could be argued that+(the)	14	***it is clear that***	11
		it is clear that	8	*it is not clear*	4
		it is estimated that	4	***it is possible to***	6
		it is possible to	13	*seems to have been*	6
		it would have been	5	*the fact that this*	4
		the fact that the	8	*there is evidence that*	5
		the fact that they	4	***to the fact that***	5
		there is no evidence	4	*was not so much*	4
		would have to be	6	*whether or not to*	5
		would need to be	4	*would be difficult to*	5
				would have to be	8
type	5	**type**	15	**type**	16
token	24	**token**	96	**token**	94

Take the commonly used hedging devices in academic writing for example. The frame 'copula *BE* + *likely to*' is frequently used in native writing to mitigate the proposition with a few variations such as *is likely to be, are likely to be, are more likely to*. In addition to this frame, native writers also appear to be able to flexibly employ other hedging devices like the 'anticipatory *it* + adjective fragment' frame (*it is clear that, it is not clear, it is possible to*), modal verbs (*would have to be, would need to be, would be difficult to*), hedging verbs (*seems to have been, it has been suggested, it can/could be argued, it is estimated that*), and hedging nouns (*there is no evidence, there is evidence that, the fact that the*, etc.) to qualify their propositions.

- This change indicates that two relatively dissimilar clusters have been merged and that the number of clusters prior to this merger ***is likely to be*** the most appropriate. (FLOB-J)
- In any case, ***it is not clear that*** such a group would comprise and

effective control if, as both Shaffer and Inhoff et al. suggest, typists, in general, exhibit a characteristic reading style. (FLOB-J)

- If a belief purports to be true for every person, it cannot be true for only some; such a belief ***would have to be*** either true for everyone, or else true for no one whatsoever. (FLOB-J)
- If activists actions are justified ***it could be argued that*** firms should withdraw from the market because they are acting unethically. (BAWE-EN)
- This does not however remove grounds for rational doubt, as though the postulation of God may be simpler than the complex answers science may give, ***there is no evidence*** to show that it is in fact more likely. (BAWE-EN)

By contrast, there are only five epistemic bundles in L2 writing: *are more likely to*, *is considered to be*, *it has been suggested, it is believed that*, and *to the fact that*. The mitigating power of these expressions used by learners, overall, seems to be weaker than those used by the native speakers above.

- All in all, obesity ***is considered to be*** a disorder in industrial countries. (BAWE-CH)
- ***It has been suggested*** that some form of impropriety linked to the use of the corporate structure that goes beyond a subjective assessment of what is "just" on any particular set of facts. (BAWE-CH)
- ***It is believed that*** this 'best practice' of ASDA has set a model for other retailers to follow. (BAWE-CH)

Apparently L1 Chinese learners of L2 English have showed some control of this specific feature in academic writing but have not demonstrated it as diversely and robustly as native writers do. Certainly it is possible that L2 learners resort to other types of expressions which may not be frequently used by native speakers for the purpose of hedging and thus

cannot be easily recognised from merely the bundle repertoire. However, a preliminary inspection like this has demonstrated how a contrastive analysis of lexical bundles lends itself to revealing the essential differences between native writing and learner writing. This marked feature in L2 learner writing, i.e. the poverty of hedging devices, will be discussed in more detail in Chapter 7.

With respect to the attitudinal/modality bundles, the bundle structures employed in the three groups of writers seem very similar (see Table 5-32). The two frames 'anticipatory *it* + adjective fragment' (*it is difficult to*, *it is important to*, etc.) and 'Adjective + *to*-clause fragments' (*must be able to*, *to be able to*, etc.) are generally used by the writers to convey obligation, directive, or ability, and the only four attitudinal/modality bundles in FLOB-J are all shared by both the student corpora.

Table 5-32 Stance attitudinal/modality bundles in Modular Study 1

BAWE-CH	Freq	**BAWE-EN**	Freq	**FLOB-J**	Freq
it has to be	4	***it is difficult to***	6	***it is difficult to***	14
it is difficult to	9	***it is important to***	17	***it is important to***	5
it is easy for	4	***it is necessary to***	7	***it is necessary to***	6
it is easy to	6	*not be able to*	6	***to be able to***	6
it is important to	4	*should be able to*	4		
it is necessary to	8	*should be placed on*	4		
must be able to	4	*that need to be*	4		
to be able to	4	***to be able to***	8		
		will be able to	4		
		would be able to	5		
type	8	**type**	10	**type**	4
token	43	**token**	65	**token**	31

5.4.4.3 Discourse Organisers

As indicated in Figure 5-8, both British students and Chinese students appear to use a fairly high number of different discourse organisers in their writing when compared to FLOB-J.

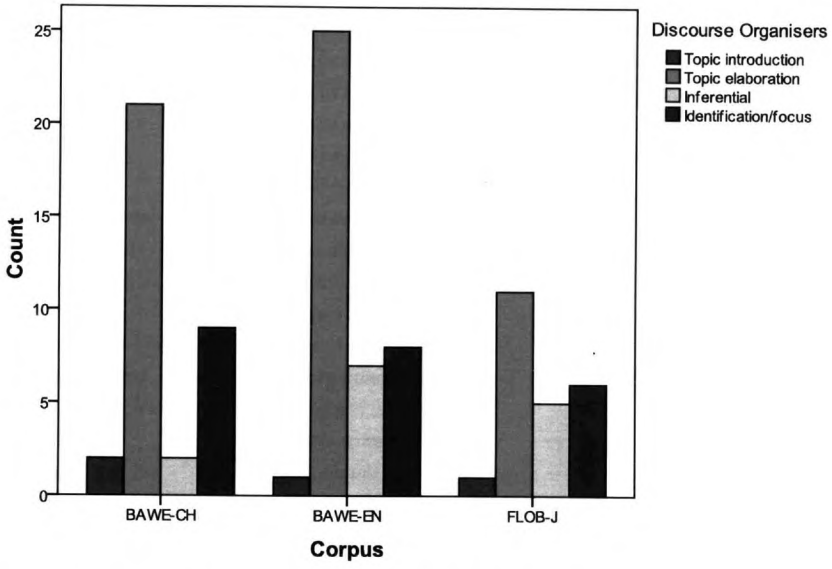


Figure 5-8 Breakdown of discourse organisers in Modular Study 1 (types)

Both groups of student writing particularly used many more discourse organisers to elaborate and/or clarify a topic. Here the different topic elaborators/clarifiers are listed with their corresponding frequencies in Table 5-33. As can be seen, native academics not only used a much narrower range of discourse organisers but also used them with lower frequencies, which might be because they have other devices other than explicit discourse markers to organise the text (e.g. possibly implicit semantic relationship). The contrasting styles in terms of the use of discourse organisers between expert writing and student writing will be further illustrated below.

Table 5-33 Topic elaboration/clarification bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
as long as the	6	*+be seen as a*	5	***an example of this***	4
as soon as the	4	***an example of this+(is)***	8	***be taken into account***	5
as well as the	16	*as well as the*	10	***can be used to***	6
can be divided into	4	*be included in the*	4	*in contrast to the*	4
can be explained by	7	***be taken into account***	5	*in more detail in*	4
can be regarded as	4	*can also be used+(to)*	5	*is concerned with the*	4
can be used to	10	*can be applied to+(the)*	7	*it has not been*	4
in addition to the	4	*can be seen as+*	5	*it is not always*	4
in order to achieve	8	*can be used for*	5	*on the one hand*	8
in order to avoid	7	***can be used to***	19	***on the other hand***	19
in order to be	5	*could be seen as*	5	*was followed by a*	4
in order to maintain	4	*could be used to*	6		
in order to make	4	*in an attempt to*	4		
in order to understand	4	***in order to make***	8		
is not only a	4	*in order to minimise*	4		
meet the requirement of	4	*is an example of+(a)*	6		
on the other hand	36	***on the other hand***	4		
pay more attention to	4	*taking into account the*	4		
that is to say	6	***that is to say***	4		
this means that the	4	***this means that the***	4		
to ensure that the	4	*to be added to*	4		
		to cope with the	4		
		to enable them to	4		
		to take into account	4		
		will be used to	4		
type	21	**type**	25	**type**	11
token	149	**token**	142	**token**	66

As can be seen, the majority of topic elaboration/clarification bundles are verb-based bundles such as 'Passive verb + prepositional phrase fragment' (*can be explained by*, *can be regarded as*, *be included in the*, etc.), 'Verb phrase + *to*-clause fragment' (*in order to make*, *to take into account*, etc.), and 'Pronoun/noun + BE/verb phrase' (*this means that the*, *that is to say*).

- `The successful stories of East Asian states` ***`can be explained by`*** `several ways including geography, history, culture and role of states. (BAWE-CH)`

- An example ***can be used to*** clarify the theory. (BAWE-CH)
- According to the report by Mintel International Group Limited (2004), consumers spending on cars grew in year 2001 excess of 10% due to the drop in prices as a result of pressure from the government. ***This means that the*** industry do compete on pricing among other things. (BAWE-CH)
- A study on domestic tourism by National Council of Applied Economics Research during 2002-2003 pointed out that nearly two third of all tourists in India traveled for social purpose (Social-cultural Drivers); ***that is to say***, traveling for social purposes, overall, stands the largest percentage of trips across the country … (BAWE-CH)

An impression of the instances incorporated with the discourse elaborators/clarifiers in Chinese student writing above is that they all seem to be verbose to a certain extent. The most noticeable example of tautology might be the last one, which repeatedly talked about travelling for social purposes in India with paraphrases. A contrast of the use with the same expression *that is to say* in professional academic writing below demonstrates one of the differences with learner writing. By use of the expression *that is to say*, professional writers did not simply repeat what has been said as learners did but progressed further by other means, e.g. giving a specific example to illustrate the prior proposition.

- It is now accepted on all sides that Britain needs more of its workforce to be vocationally trained to intermediate levels; ***that is to say***, to craft or technician standards as represented, for example, by City and Guilds examinations (at part 2) or BTEC National Certificates and Diplomas. (FLOB-J)

Among the discourse elaborators/clarifiers, two kinds of frames are found to be prevalently used in both British student writing and Chinese student writing: '*in order to* + verb' and '(*can/could*)+*be seen/regarded as*+(*a*)'. Overall, Chinese students used more word combinations incorporated with '*in order to*' while British students used a slightly smaller

number of such word combinations, but the frequencies in both groups of student writing are a lot higher than that in professional academic writers, which has been discussed in Section 5.3.4.2. Chinese students used a variety of verbs following '*in order to*' in defined lexical bundles: *achieve, avoid, be, maintain, make*, and *understand* while there are two verbs that fit into this bundle frame in British student writing: *make* and *minimise*. While native professional writers also used diverse verbs with this frame, none of the frequencies reached the threshold of four times.

Some interesting findings also emerge from a further examination of collocates with the frame '*be seen/regarded as*' (see Table 5-34). It appears that native professional writers collocate diverse modal auxiliaries such as *may, must, would, should, can*, and *could* with '*be seen/regarded as*', but none of them meet the cut-off frequency requirement, yet students' preference for the modal auxiliaries *can* and *could* collocated with '*be seen/regarded as*' bring these word sequences forward to pass the required frequency threshold. It also has to be mentioned that since the frame '*be seen/regarded as*' mostly collocates with modal auxiliaries, it means that these bundles usually entail some element of epistemic stance despite being assigned under the category of discourse organisers. This type of bundles is usually used to express an alternative perspective for a preceding proposition.

- As a company they are known for their cheaper prices compared to some competitors, but in the market low prices ***can be seen as*** a signal to start a price war and will inevitably be under cutting competitors like estate agents.(BAWE-EN)
- Everyone has his/her own personal preferences in processing information and solving problems. These personal preferences ***can be regarded as*** different learning styles. (BAWE-CH)

Table 5-34 Frequency of *be seen as/be regarded as* in Modular Study 1

Frequency	BAWE-CH	BAWE-EN	FLOB-J
be seen as	72	19	7
be regarded as	137	1	13

The rest of subcategories of discourse organisers, i.e. topic introduction bundles, identification/focus bundles, and inferential bundles, are presented below (Table 5-35, Table 5-36, and Table 5-37). As can be seen, many of them in two student corpora are composed of a verb element and thus categorised as VP-based bundles. In the next section, we will turn to see whether there is any interaction between structural and functional categorisation, thereby leading to the distinctive patterns contrasting native expert writing represented in FLOB-J with student writing represented in BAWE.

Table 5-35 Topic introduction bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
essay is going to	4	*in this essay I*	4	*in the first place*	6
last but not least	5				
type	2	**type**	1	**type**	1
token	9	**token**	4	**token**	6

Table 5-36 Identification/focus bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
as one of the	8	***is one of the***	12	***it can be seen that***	5
it can be seen	4	***it can be seen+(that)***	12	***one of the most***	10
one of the most	11	*it should be noted*	4	***that there is a***	9
(played)+an important role in	5	*one of the main*	11	*that there is no*	7
		one of the most+(important)	7	*to that of the*	4
bear in mind that	5	*than that of the*	7	***was one of the***	4
is one of the	9	*there would be no*	5		
that there is a/an	17	***was one of the***	6		
we can see that	7				
will focus on the	4				
type	9	**type**	8	**type**	6
token	70	**token**	64	**token**	39

Table 5-37 Inferential bundles in Modular Study 1

BAWE-CH	Freq	BAWE-EN	Freq	FLOB-J	Freq
as a result of	12	*(due)+to the fact that*	8	***as a result of***	9
this is due to	4	*and as a result*	7	*in the light of*	6
		as a result of+(the)	17	*in the sense that*	8
		because it is not	4	*in view of the*	4
		this is due to+(the)	5	*the results of the*	4
		this may be due+to	4		
		to a lack of	5		
type	2	**type**	7	**type**	5
token	16	**token**	50	**token**	31

5.5 Relationship between Structural and Functional Categorisation

Biber et al. (2004, pp. 397-398) found a strong interaction between bundle structures and bundle functions. The association between form and function is consolidated to a very large extent in the corpora of academic writing investigated in this project. As Figure 5-9 shows, referential expressions and stance bundles are two extreme examples of the sharp contrast. Noun and prepositional phrase fragments make up the majority of referential expressions (over 90%) while verb phrase fragments constitute the majority of stance bundles (around 90%). Discourse organisers seems to be the category that is most evenly distributed, with VP-based bundles comprising around 60% and NP-based and PP-based bundles making up the remaining 40%.

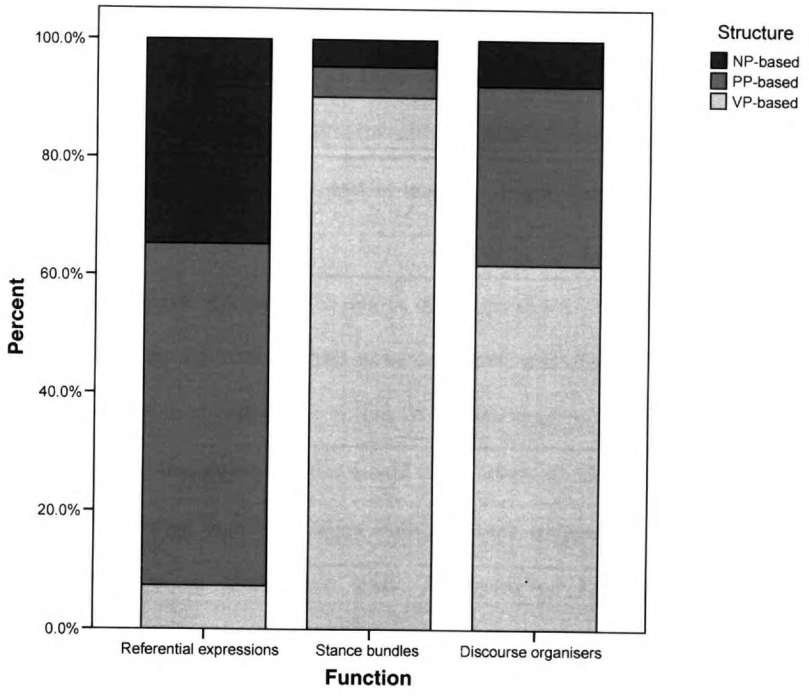


Figure 5-9 Interaction between structural and functional categorisation in Modular Study 1

This strong relationship is also illustrated by the dominance of referential bundles in native expert writing in FLOB-J as two thirds of its bundles are NP-based and PP-based bundles. By contrast, over half of the bundles found in the two groups of student writing are VP-based bundles, and accordingly there is far less reliance on the referential expressions. As we have seen, the student writers instead used more discourse organisers, nearly twice as many as the native professional writers.

This intensive interaction between form and function of lexical bundles will be attested again in Modular Study 2. As writers of different groups tend to use different types of word combinations, it will be interesting to see if this strong relationship between form and function of lexical bundles sustains across various groups of writers.

5.6 Keyness Analysis

As has been discussed in Section 4.1.3, a keyness analysis can be used to complement the previous quantitative comparison made between the three groups of writing by revealing significant 'overuse' and 'underuse' in student language when compared with the reference corpus, FLOB-J.

The *p* value, which represents the degree of danger of error, was set at 0.001. Firstly, FLOB-J was used as the reference corpus to be compared with BAWE-EN and BAWE-CH respectively because it is regarded as a corpus of written academic English and reasonably represents the standard that novice writers would try to attain. In addition, remember that the determination of a lexical bundle involves both a cut-off frequency of four times and a dispersion threshold of at least three texts. The retrieved key clusters were hence crosschecked with the list of target lexical bundles to ensure that the key clusters met the cut-off frequency and dispersion requirements set in this thesis. After removing clusters that were not within the defined lexical bundles, the filtered key bundles were finalised as shown in Table 5-38 with the key value in the brackets. A positive value indicates that the bundle in question is statistically more frequent when compared to FLOB-J while a negative value means that the bundle in question is significantly less frequent.

Table 5-38 Key lexical bundles in BAWE-EN and BAWE-CH with FLOB-J as the reference corpus

Corpus	Overuse/underuse	Key lexical bundles
BAWE-EN	Overuse	***is one of the*** (17), *it could be argued that* (17), *the use of the* (16), *the development of the* (16), *one of the main* (16), *in order to make* (12), *the fact that the* (12), *the extent to which* (12)
	Underuse	*in the context of* (-26), ***as a function of*** (-22), *on the basis of* (-20)
BAWE-CH	Overuse	*in the long run* (19), ***is one of the*** (19), *in order to achieve* (12), *as well as the* (11)
	Underuse	***as a function of*** (-22), *the way in which* (-16)

* The key bundles appearing in both BAWE-EN and BAWE-CH are indicated in bold print.

In relation to the key bundles uncovered, it appears that British students overused and underused more lexical bundles than Chinese students did, which can be employed to supplement the results of chi-square residuals (cf. Sections 5.3.3 and 5.4.3) by illustrating what lexical bundles in BAWE-EN unexpectedly contributed more to the significant difference than BAWE-CH, particularly in token distribution. Nevertheless, there is still some similarity between these two groups of student writing. For example, in comparison with FLOB-J, they both overused *is one of the* and the '*in order to* +V' pattern, and also underused *as a function of*. As a matter of fact, many of the bundles in Table 5-38 have been discussed earlier in this chapter. For example, the peculiar prevalence of *use* in British student writing is evidenced again in the overuse of *the use of the* (cf. Section 5.3.4.1). The preference of the frame '*in order to* +Verb' in student writing can be found in both student groups: *in order to make* and *in order to achieve* (cf. Sections 5.3.4.2 and 5.4.4.3).

As a whole, the statistical analysis here consolidates the earlier findings to a large extent. On the other hand, it also provides some information which might have been overlooked in the discussion earlier. First of all, the underuse of *as a function of* foregrounded by the keyness analysis might be related to the fact that FLOB-J used as the reference corpus actually contains more texts based on hard-science subjects. In addition, it appears that both

student groups used *is one of the* to identify and elaborate a proposition, with nine occurrences in BAWE-CH and twelve occurrences in BAWE-EN. In student writing, this expression often collocates with the superlative form as some sort of hedging marker to modify the proposition (e.g. *is one of the cheapest*, *is one of the most*). By contrast, the native expert writing does not show this tendency, although there is one similar bundle with the past tense *was one of the* (with only four occurrences). Lastly, it is intriguing to see that the British students used two important expressions in academic writing *in the context of* and *on the basis of* even less frequently than the Chinese students; these bundles are highly frequent in native academic writing and thus could be identified as important makers for academic/formal writing. This somewhat confirms the observation that the native student writing generally does not appear to be much closer to native expert writing in terms of the use of lexical bundles than learner student writing does, which has been evidenced by various analyses discussed in this chapter. The only exceptions seem to be the use of passive verb forms, hedging devices, and extent/degree modifiers (cf. Sections 5.4.4.1 and 5.4.4.2).

The keyness analysis was also conducted with BAWE-CH as the target corpus and BAWE-EN as the reference to see whether Chinese students exhibited any significant difference in the use of lexical bundles with British peers. The results are presented in Table 5-39. It is not surprising to see that *in the long run* pops up once more as being overused by Chinese students, as it has been discussed that both native groups of writing rarely used this idiomatic expression (cf. Section 5.4.4.1). Again, the keyness analysis between BAWE-EN and BAWE-CH may be used to explain the results of the chi-square residuals reported in Sections 5.3.3 and 5.4.3. Combined with the results in Table 5-38, now we have a clearer idea with regard to which bundles contributed to the significant difference discussed earlier.

Table 5-39 Key lexical bundles in BAWE-CH with BAWE-EN as the reference corpus

Corpus	Overuse/underuse	Key lexical bundles
BAWE-CH	Overuse	*on the other hand* (32), *that there is a* (19), *in the long run* (19), *at the same time* (15)
	Underuse	*it is possible to* (-19), *one of the main* (-17)

5.7 Summary of Findings

The results for the retrieval of lexical bundles, structural and functional distribution, and keyness analysis can be summarised as below:

- as shown in the totals columns of Table 5-10, Table 5-11, Table 5-24, and Table 5-25, the number of lexical bundles appears to increase with the advance of writing competency for both the range of lexical bundles used (types) and overall frequency of lexical bundles (tokens), which will be further discussed in Chapter 7;
- in terms of structural distribution regardless of types or tokens, NP-based and PP-based bundles altogether occupy nearly 70% of lexical bundles in FLOB-J whereas in the two groups of student writing in BAWE-EN and in BAWE-CH, roughly half of the bundles are composed of VP-based expressions;
- in terms of functional distribution regardless of types or tokens, the proportion of referential expressions in FLOB-J reaches as high as over 60% whereas both BAWE-EN and BAWE-CH exhibit a much less referential use, and instead both have discourse organisers as their most prevalent function, taking largely over 40% of the total bundles;
- chi-square tests reveal that there is always a significant difference among FLOB-J and the two groups of student writing regardless of structural or functional distribution and regardless of type or token distribution, and the difference is always more pronounced in *token* distribution than *type* distribution;

- a further examination of chi-square standardised residuals suggests that NP-based and VP-based bundles contribute greatly to the significant difference of structural distribution among the three corpora while referential expressions and discourse organisers make the major contribution to the significant difference of functional distribution;
- the British student writing generally does not appear to be quantitatively much closer to native expert writing in terms of use of lexical bundles than learner student writing does, which has been evidenced by various quantitative analyses discussed in this chapter;
- in terms of structural analysis, however, learner writing also distinguishes itself from the two groups of native writing by its lack of NPf bundles (e.g. *the way in which*) as well as by containing fewer passive-verb bundles (e.g. *can be found in*);
- in terms of functional analysis, learner student writing is also distinct from native writing with much fewer extent/degree modifiers (e.g. *the extent to which*) and hedging devices (e.g. *is likely to be*) as well as favouring certain deictic bundles (i.e. *in the long run, in the recent years,* and *all over the world*);
- there is a strong relationship between structural and functional categorisation, particularly in referential expressions (mostly composed of NP-based and PP-based bundles) and stance bundles (mostly composed of VP-based bundles); and
- the keyness analysis consolidates many of the findings discussed earlier (e.g. overuse of the '*in order to* +V' frame), and, on the other hand, it also reveals some of the patterns which have been overlooked earlier (e.g. overuse of *is one of the* and underuse of *as a function of*).

The discussion in this section set out from the structural and functional distribution of lexical bundles with the aim of disclosing the difference and/or similarity of phraseological aspect between the three corpora. A deeper investigation, however, suggested that the

quantitative analysis of lexical bundles needed to be complemented and supported by other measures such as keyness analysis, type/token ratios, and concordance checks. By taking advantage of such a hybrid methodology, a number of distinctive features varying with levels of writing competency have been unveiled. In the next chapter, we shall see whether learner writing across two different proficiency levels will also reflect the similar developmental patterns revealed here.

Chapter 6 Modular Study 2: Lexical Bundles across Learner Proficiencies

In this chapter, lexical bundles in L2 learner writing across two proficiency levels will be discussed. This chapter is distinguished from the previous chapter in that essays selected from the Longman Learners' Corpus were investigated after the proficiency had been determined by a robust rating procedure as generally adopted in large-scale tests. This way of proficiency determination, to my knowledge, has rarely been reported in second language research except for the studies which used readily available examinee performance data from language proficiency tests such as IELTS (cf. Section 2.3).

The first part of this chapter attempts to answer the third methodological/procedural research question (see Section 3.1) by describing the rating process in detail, starting from benchmarking to analysis of ratings. On the basis of rating results and corpus comparability, two small learner subcorpora representing different proficiency levels were finalised. The second part of this chapter then deals with the first and third analytical research questions, i.e. investigating the lexical bundles in these two subcorpora in order to reveal the developmental pattern in the written performance of L1 Chinese learners of L2 English. Following the same analytical framework adopted in the previous chapter, the structural and functional distributions of lexical bundles will be discussed in terms of quantitative and qualitative perspectives. A keyness analysis will also be conducted. It is expected that a similar developmental pattern as disclosed between expert writing and student writing will be found (e.g. more proficient writing would contain more NP-based and PP-based bundles in terms of structures and more referential bundles in terms of functions), and we shall see whether this assumption would hold true in this chapter.

6.1 Determination of Proficiency

As discussed in Section 2.3, this thesis argues that a rating procedure as generally adopted in high-stakes language tests, instead of extra-linguistic judgement (such as years of learning), should be utilised to determine learner proficiency in second language research before any valid claims can be made. The procedure of standardisation of judgement used in this thesis originated from *the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2003).

Six band levels are distinguished in the Common European Framework of Reference (CEFR), from the lowest level A2 to the most advanced level C2 (for the CEFR global descriptors, see Appendix 2). Holistic scoring would be adopted with the use of a rating scale from the Manual which consists of overall descriptors as well as three analytical criteria: range, coherence, and accuracy (for the CEFR scale, see Appendix 3 CEFR Written Assessment Criteria Grid). Working with experienced raters, I set off from the familiarisation training and continued with benchmarking. Finally, approximately 1,000 learner essays from the Longman Learners' Corpus were marked by at least two raters. Various statistical measures, such as descriptive statistics and Multi-faceted Rasch analysis, were then carried out and will be discussed below. The statistical measures not only guaranteed the quality of rating but also greatly contributed to selection of data for the follow-up investigation. The whole process can be summarised in Figure 6-1 below and will be described in detail in the next section.

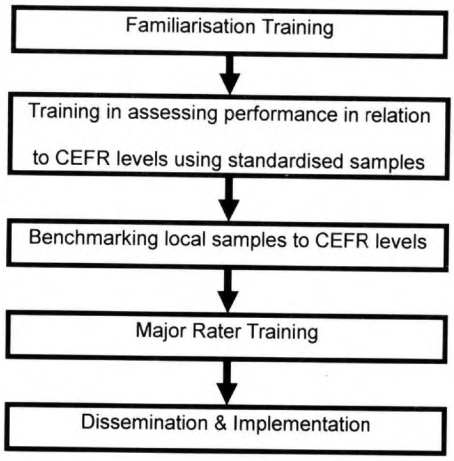


Figure 6-1 Visual representation of standardisation of judgments (extracted and modified from Figure 1.1 Section A in *Reference supplement to the preliminary pilot version of the Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2004))

6.1.1 Benchmarking

6.1.1.1 Procedures

Five members in the Language Testing Research Group at Lancaster University participated in the benchmarking. The invited testing professionals, all doctoral students in Linguistics and English Language, were given a standardisation package which contained detailed instructions of benchmarking procedures, the CEFR rating scale for writing assessment, a number of standardised samples and practice scripts, and twenty local performance scripts. The participants were overall fairly familiar with the Common European Framework of Reference (CEFR) in the light of their research interests, and yet they were advised to again thoroughly familiarise themselves with the CEFR level descriptors before marking the scripts.

The local performance scripts were first selected by the researcher from the Longman Learners' Corpus (LLC), all of which were produced by L1 Chinese learners of L2 English.

Then these scripts were marked by ten members of the same Language Testing Research Group in a pre-benchmarking session. On the basis of band levels assigned to the scripts in pre-benchmarking, these learner essays were again reviewed and chosen by the researcher afterwards to ensure that the whole set of twenty essays for benchmarking covered a sufficiently wide range of CEFR levels.

Given the difficulty in organising a one-day benchmarking workshop considering the participants' tight schedules, the standardisation scheme was devised as the following three stages which allowed more flexibility through self-training and marking in the raters' own time. The five raters then gathered at the standardisation meeting to discuss their marking.

- **Stage 1: Familiarisation**

At this preparation stage, the participants read eight samples and practiced rating three scripts, which were retrieved from the sample papers of the Cambridge ESOL Main Suite (Cambridge ESOL, 2007) which have been standardised to the CEFR levels[28], so as to acquire experience in relating written descriptors to particular levels of performance. Five exams of different levels are available in the Cambridge Main Suite: KET, PET, FCE, CPE, and CAE[29], and each of them have been linked to a corresponding CEFR level: A2, B1, B2, C1, and C2. The General Mark Scheme for writing awards six band scores in each exam: Band 0 to Band 5. The primary selection criterion of scripts for the current study is the medial

---

[28] The Council of Europe released a set of written samples officially calibrated with the CEFR in 2008 (http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html) while the benchmarking in this project was completed in November, 2007. The pilot written exemplars used in the official standardisation set, however, were also provided by Cambridge ESOL from Cambridge Main Suite of certificated examinations.

[29] The Cambridge Main Suite of exams is composed of the following: KET (Key English Test), PET (Preliminary English Test), FCE (First Certificate in English), CAE (Certificate in Advanced English), and CPE (Certificate of Proficiency in English).

Band 3, also the pass threshold from the Cambridge exams, with a few exceptions of scripts that are awarded Band 4. According to the exam handbook, those scripts that are awarded Band 3 demonstrate adequate performance at that level. That is to say, they are neither too superior (e.g. Band 5) nor too inferior (e.g. Band 2) and should be appropriate for the purpose of rater training.

The raters were not asked to form a consensus on the level of the standardised scripts at the current stage but rather to arrive at the recommended CEFR level by applying the criteria in the rating scale. They were also encouraged to note down any questions they might have come across, if any, which they could choose to discuss with the researcher or at the face-to-face discussions later on.

- **Stage 2: Benchmarking Local Samples**

At this stage, the raters were asked to apply their judgment formed at the familiarisation stage to the assessment of local performance samples, in this case the LLC samples. They marked 20 scripts in their own time and sent the results to the researcher so that the rating statistical report could be generated before the standardisation meeting. This allowed the discussion to focus on the scripts that appeared problematic in the statistical report.

- **Stage 3: Standardisation Meeting**

A standardisation meeting took place after all the raters submitted their scores. At the meeting, five raters discussed the questions that were brought forth from the previous stages, disagreement of marks given, and distinctions between levels. Other issues concerning application of the CEFR scale and the rating statistics were also raised and discussed, which will be further addressed below.

6.1.1.2 Statistics for Benchmarking

A statistical report of the 20 ratings given by the five raters was prepared including information of descriptive statistics, inter-rater reliability and Multi-faceted analysis. The six

CEFR band scores were converted into numerals in order to enable statistical procedures to be carried out. For instance, C2 was transcribed into 6, C1 into 5, and so on (see Table 6-1). Although no A1 scripts were included in the training materials, an A1 level was still included in the statistical scheme in case an A1 was awarded (in benchmarking or the follow-up major rating). In addition, pseudonyms were used hereafter to conceal the identity of all the raters.

Table 6-1 Transcription of CEFR levels to numerals

C2	6
C1	5
B2	4
B1	3
A2	2
A1	1

The general descriptive statistics are presented in Table 6-2. The standard deviation and the mean are two conventional measures used frequently for reporting the results of rating. The former indicates the dispersion of scores awarded by the rater while the latter displays the extent of rater harshness/leniency. By comparing the mean values, it is then found that Shawn is the strictest rater with a mean of 3.45 while Cheryl and John are the most lenient raters with a mean of 4.00. Standard deviations range from .826 to 1.257 with Shawn as the rater who tended to mark within a narrow range of band scores and Cheryl and John as the raters whose ratings spread out the most. In addition, Kat and Shawn did not award any scripts the highest level of 6 (C2) while the other three raters did.

Table 6-2 Descriptive statistics for benchmarking

	Mean	Std. Deviation	N	Minimum	Maximum
Kat	3.65	1.089	20	2	5
Shawn	3.45	.826	20	2	5
Mike	3.80	1.152	20	2	6
Cheryl	4.00	1.257	20	2	6
John	4.00	1.257	20	2	6

Various statistical measures can be used to calculate inter-rater reliability, which refers to the extent of agreement between two or more raters with regard to their judgment of candidates' performance. Yet it is generally recognised that each of these measures has its own constraint (Jonsson & Svingby, 2007). First, the non-parametric Spearman's correlation coefficient *rho* was calculated. As Table 6-3 indicates, the benchmarking overall reveals significantly strong correlations between any two raters, ranging from .741 to .932 ($p<0.05$, two-tailed). We can see that Kat and John show the highest correlation (rho=.932, $p<0.05$, two-tailed). It is of note that they are the most experienced raters within the five benchmarking raters. Kat used to work for a language centre in Taipei, an institution in charge of various language proficiency tests, and her responsibility included raters' training and monitoring raters' performance. John also has a great deal of experience as a rater in speaking and writing tests in the context of British higher education.

Table 6-3 Correlation coefficients for benchmarking

Spearmans' correlation coefficients (n=20) Mean=.83, Max= .932, Min= .741					
	Kat	Shawn	Mike	Cheryl	John
Kat	1	.845(**)	.815(**)	.830(**)	.932(**)
Shawn	.845(**)	1	.743(**)	.818(**)	.800(**)
Mike	.815(**)	.743(**)	1	.741(**)	.853(**)
Cheryl	.830(**)	.818(**)	.741(**)	1	.927(**)
John	.932(**)	.800(**)	.853(**)	.927(**)	1
Cronbach's *alpha* index = .957					

** Correlation is significant at the 0.05 level (2-tailed).

Cronbach's *alpha* was also calculated for internal consistency (cf. Kaftandjieva & Takala, 2002, p. 112). This inter-rater reliability index is encouragingly high at the level of .957. This suggests that the constructs the raters applied in marking LLC essays with the CEFR scale were very similar. Spearman's correlation coefficients as well as Cronbach's

alpha, however, cannot thoroughly address the issue of rating reliability because they do not reflect the true difference between raters. Take Spearman's correlation coefficients for example. It indicates the strength of association of ratings given by individual raters as opposed to the actual scoring. If one rater makes use of four CEFR Levels B1 to C2 (four continuous levels: B1, B2, C1, and C2) and the other rater tends to give another four continuous CEFR levels A2 to C1 (A2, B1, B2, and C1), they can still have a high coefficient despite a gap of one CEFR level between the scores awarded by the two raters. Most important of all, neither inter-rater reliability indexes nor descriptive statistics can provide information as to adjustment of ratings in view of the extent of rater disagreement even though it has been identified. Next, we shall see how the Many-faceted Rasch analysis can remedy for the above problems and calculate the candidates' ability estimates taking into account the variable of raters.

The Multi-faceted Rasch measurement is perhaps the only powerful model to date which copes with psychological measurement in the human sciences (Bond & Fox, 2007). It offers 'useful approximations of measures that help us understand the processes underlying the reason why people and items behave in a particular way' (*ibid*, p. 8). Rasch measurement has been widely applied in diverse areas, particularly in psychological or educational measurement such as patient assessment and language testing (e.g. Lumley & McNamara, 1995; Schaefer, 2008; Van Moere, 2006, to name just a few). In the case of essay grading, it can be utilised to gauge the abilities of candidates (Longman learner writers in this case) after factoring in different facets such as raters, conditions (e.g. whether timed or not) or instrument (e.g. tasks). As the variable of writing tasks is not well controlled in the Longman Learners' Corpus, despite being documented, there are only two facets involved in the Rasch analysis here: raters and candidates. The software *FACETS* (Linacre, 2008) was utilised to compute Multi-faceted Rasch analysis with the Longman rating data. A selection of statistical

figures generated by FACETS which facilitate the determination of final CEFR level assignment will be illustrated below, and how these statistics contribute to the understanding of rater and learner performance will be discussed, too.

One basic principle behind the Rasch Model is that it searches for 'patterns' in the data so as to make the behaviour of facets such as candidates or raters predictable (McNamara, 1996). The *fit* statistics, one of the key concepts embodied in the search for patterns in the Rasch modelling, 'summarize for each rater, item, person, etc. the extent to which the difference between expected and observed values are within a normal range' (*ibid*, pp. 141-142). Take the raters for example. For the presence of a good fit value, it means that the scores given by the rater in question match the general pattern. A 'misfit' refers to a condition in which the scores of the rater do not conform to the patterns found. By contrast, if the scoring of the rater lacks normal variability, i.e. being too predictable, then this is a case of 'overfit'.

Table 6-4 is a logit scale, which exhibits the spread of candidate ability, rater severity, and the CEFR scale in the same visual representation constructed by the Rasch Model with around 14 logits (+8 to -6). Logit is the unit of measurement used in Rasch measurement referring to the probability of a certain event and is expressed in the far left column (*Measr*). The top of the distribution represents greater candidate ability and greater rater severity, the format of which is adjustable in the specifications before running the analysis. As can be seen, #19 is rated as the most able candidate while #1 the least able candidate. In terms of raters, Shawn is the most severe rater and Cheryl and John the most lenient raters. The far right column (*Scale*) is the numerical representation of the CEFR scale, i.e. 2 to 6 for levels A2 to C2. Because the purpose of this benchmarking is to select scripts representative of each level, the scripts used were carefully selected to ensure that they cover a wide range of abilities as has been discussed earlier. We can thus see that 20 candidates distribute from 2 (A2) to 5 (C1)

as what was expected, but no candidate is positioned at the highest level 6 (C2) or the lowest level 1 (A1). The ablest candidate #19 is at the borderline of 5 (C1) approaching 6 (C2). The lowest level 1 (A1) does not appear in this graphic representation since no one gave any A1 ratings. As to rater severity, the raters appear to cluster around ±1.0 logit except for Shawn. Whether these values fall within an acceptable range will be further discussed later.

Table 6-4 All Facet vertical 'rulers' for benchmarking

Measr	+candidate	-rater	Scale
+ 8 +		+	+ (6) +
|	|	|	| --- |
+ 7 +	#19	+	+ +
|	| #16 #17 #9	|	| |
+ 6 +		+	+ 5 +
|	|	|	| |
+ 5 +		+	+ +
|	| #18 #7	|	| --- |
+ 4 +		+	+ +
|	|	|	| |
+ 3 +		+	+ +
|	| #14	|	| |
+ 2 +		+	+ +
|	|	| Shawn	| |
+ 1 +	#10	+	+ 4 +
|	|	| Kat	| |
* 0 *		*	Mike * *
|	| #15	|	| |
+ -1 +		+	Cheryl John + +
|	| #11 #13 #4 #5 #6	|	| |
+ -2 +		+	+ --- +
|	|	|	| |
+ -3 +		+	+ +
|	|	|	| |
+ -4 +	#2	+	+ +
|	|	|	| |
+ -5 +		+	+ +
|	| #8	|	| |
+ -6 +		+	+ 3 +
|	|	|	| |
+ -7 +		+	+ +
|	|	|	| |
+ -8 +		+	+ +
|	|	|	| |
+ -9 +	#3	+	+ +
|	|	|	| --- |
+ -10 +	#12	+	+ +
|	|	|	| |
+ -11 +	#20	+	+ +
|	|	|	| |
+ -12 +	#1	+	+ (2) +
|Measr|	+candidate	-rater	Scale|

Table 6-5 is a candidate measurement report. The far right column indicates the number of candidate (#1, #20, #12...). The third column from the left, *Obsvd Average*, presents the average raw scores awarded to each script. The fourth column, *Fair-M Avrage*, reveals the band scores calibrated by the Rasch system after factoring in the variable of raters. The differences between raw average scores and Rasch-scaled scores, as can be seen, are minimal. The next column, *Measure*, is the estimates of candidate ability expressed with logits, as shown in the column of candidate in Table 6-4. The *fit* statistics, particularly the *infit* values, are generally considered to be the most informative figures and thus worth the most attention in this analysis. They represent the extent to which the observed scores differ from what the system expects the patterns to be. As the generally acceptable range of *Infit MnSq* (infit mean square)[30] falls within 0.75 to 1.3 (McNamara, 1996, p. 173), #20, #12, and #15 can be immediately spotted as the 'misfitting' candidates, having values greater than 1.3. A misfit, indicated by an over-high fit value, suggests that the variation of band scores awarded to a particular candidate exceeds the acceptable range predicted by the model. In terms of 'overfit' with extremely low fit values, the model suggests that the ratings might lack variation because they overly fit the patterning by being too consistent. The traditional school would tend to aim for overfitting rating with an attempt to eliminate misfitting rating. The Rasch model, on the other hand, accepts variability as part of human behaviour, and consequently, overfits could be undesirable on the ground that they could not provide sufficient information for the facet investigated. In the case of test items (e.g. reading comprehension tests), overfits suggest a 'lack of independence of items' (*ibid*, p. 172), and thus they could be regarded as being redundant because of not making contribution to the understanding of candidate ability. As for the LLC benchmarking discussed here, #1, #8, and

[30] The Infit statistics, with the notions of 'misfit' and 'overfit', have different implications to various facets (e.g. raters, or candidates), as will be discussed in the rest of this section.

#10, with the three lowest Infit MnSq values, are the scripts that received the same five scores from all of the raters. As the purpose of benchmarking is to seek exemplar scripts that are representative of each level, i.e. standard setting, overfits do not seem to be an issue. Instead, the overfitting scripts appear to satisfy the need and hence are appropriate to be selected as sample essays for raters' training materials. By contrast, the misfitting scripts, which arouse great variability in ratings, could be a concern for standard-setting and therefore would need to be examined carefully.

Table 6-5 Candidate measurement report for benchmarking

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	candidate
10	5	2.0	2.04	(-12.52	1.90)	Minimum					1	#1
11	5	2.2	2.16	-11.06	1.17	1.68	1.1	5.49	2.3	-.65	20	#20
12	5	2.4	2.38	-9.91	1.00	1.57	1.3	2.31	1.8	-.72	12	#12
13	5	2.6	2.63	-8.89	1.02	.49	-1.1	.44	-1.0	1.89	3	#3
15	5	3.0	3.01	-5.62	1.65	.06	-.7	.05	-.7	1.36	8	#8
16	5	3.2	3.16	-3.75	1.14	1.24	.5	1.14	.4	.80	2	#2
18	5	3.6	3.62	-1.66	1.01	.48	-1.1	.43	-1.1	1.82	4	#4
18	5	3.6	3.62	-1.66	1.01	.48	-1.1	.43	-1.1	1.82	5	#5
18	5	3.6	3.62	-1.66	1.01	.83	-.2	.72	-.3	1.33	6	#6
18	5	3.6	3.62	-1.66	1.01	1.21	.5	1.34	.7	.65	11	#11
18	5	3.6	3.62	-1.66	1.01	.83	-.2	.72	-.3	1.33	13	#13
19	5	3.8	3.84	-.51	1.16	1.41	.7	1.35	.6	.71	15	#15
20	5	4.0	4.00	1.16	1.43	.08	-.9	.06	-.9	1.40	10	#10
21	5	4.2	4.15	2.75	1.09	1.15	.4	1.13	.4	.90	14	#14
23	5	4.6	4.55	4.47	.83	1.18	.4	1.03	.2	.98	7	#7
23	5	4.6	4.55	4.47	.83	.99	.1	.91	.0	.99	18	#18
26	5	5.2	5.20	6.35	.79	.43	-1.1	.45	-1.1	1.78	9	#9
26	5	5.2	5.20	6.35	.79	.93	.0	.93	.0	1.16	16	#16
26	5	5.2	5.20	6.35	.79	.55	-.8	.56	-.8	1.81	17	#17
27	5	5.4	5.42	7.01	.83	.30	-1.5	.32	-1.4	1.94	19	#19

As has been discussed, the misfitting scripts with high Infit MnSq values are the ones with too great variation in the ratings received to the extent that the Infit MnSq values have gone beyond the acceptable range expected by the model. Take #12, with a high Infit MnSq of 1.57, for example (see Table 6-5). Remember that both descriptive statistics and Facet vertical rulers have shown that Shawn is the strictest rater while Cheryl and John are the most lenient raters. As can be seen in Table 6-6, however, this pattern has been violated as a result

of an unusually higher score B1 given by Shawn and another unusually lower score A2 given by John. There could be various explanations. For one, when Shawn and John marked this particular script, they might have taken into account factors which could not be possibly known simply from the ratings given. For another, it is also possible that Shawn and John are simply not very consistent raters. Further discussions about rater consistency in the following rater measurement report may help clarify these issues.

Table 6-6 Instance of misfitting #12 in benchmarking

#	Kat	Shawn	Mike	Cheryl	John
12	A2	B1	A2	B1	A2

With regard to the facet of raters, two key concepts are important in rating: rater severity and rater consistency. Similarly with the candidate measurement report, rater performance can be represented in Table 6-7 produced by FACETS. Rater severity is expressed with logits in Column 5 (*Measure*). Van Moere (2006, pp. 424-425) used a range of ±1.00 logits as an indicator for over leniency (<-1.00 logit) or over severity (>1.00 logit). Following this criterion, the rater Shawn with a measure of 1.61 logits would be labelled as too harsh a rater, although we have to bear in mind that there is still no universally accepted range as to examiner severity (Develle, 2008). Relative consistency, the other important rater characteristic, can be observed from *Infit MnSq* (mean square). Again, a rule-of-thumb acceptable range for Infit MnSq values proposed by McNamara (1996) is 0.75 to 1.3. Similarly, Wright and Linacre (1994) also suggested a range of 0.6 to 1.4 for an acceptable consistency level. Low fit values indicate that there is less variation than the pattern predicts as can be seen in John 0.46 and Kat 0.53, both of which are fairly low in comparison with other raters. McNamara (1996, pp. 139-140) mentioned two interpretations for low fit values: either they are very consistent raters or they use only a limited range of the scale. However,

descriptive statistics shown in Table 6-2 do not support the latter hypothesis as Kat had a reasonably high standard deviation while John had the highest standard deviation. In addition, as discussed in the section about inter-rater reliability (cf. Table 6-3), these two raters both have been involved in professional marking for a substantial period of time. Given the above facts, it is then more reasonable to conclude that John and Kat are highly consistent in marking. The reason why their Infit MnSq values are much lower than the recommended range (0.6 or 0.75) might be due to the small samples marked in this benchmarking (n=20). As for the other three raters, the Infit statistics show that they all fall within the acceptable range, and hence there is no need for retraining or elimination of any raters from the statistical report for standard-setting.

Table 6-7 Rater measurement report for benchmarking

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	N	rater
78	19	4.1	3.96	-1.05	.51	1.07	.3	.86	-.2	1.10	50.0	54.6	4	Cheryl
78	19	4.1	3.96	-1.05	.51	.46	-2.0	.40	-1.8	1.61	57.5	54.6	5	John
74	19	3.9	3.86	-.10	.48	1.16	.6	1.33	.8	.61	47.5	57.4	3	Mike
71	19	3.7	3.75	.59	.49	.53	-1.8	.48	-1.4	1.58	57.5	56.2	1	Kat
67	19	3.5	3.51	1.61	.53	.96	.0	2.14	1.8	.84	45.0	50.6	2	Shawn

The last statistic generated by FACETS worth reporting is the unexpected responses as shown by Table 6-8, which points out that Shawn's marking of #20 exhibited an outlying observation which deviates far from expectation. From the previous discussions, we have seen that Shawn is the harshest judge among the five raters. However, as Table 6-9 shows, while the other four raters all agreed on the band level of A2, Shawn gave an unusually (especially for him) higher score of B1, which by no means matches the pattern predicted by the model.

Table 6-8 Unexpected responses generated by FACETS

Cat	Step	Exp.	Resd	StRes	Nu can N rater
3	3	2.0	1.0	5	20 #20 2 Shawn

Table 6-9 Unexpected marking: #20

#	Kat	Shawn	Mike	Cheryl	John
20	A2	B1	A2	A2	A2

As a matter of fact, #12, #15, and #20 have been defined as misfitting scripts in the candidate measurement report (Table 6-5). If #20 is compared with the other misfitting items, say #12 (see Table 6-6), more than one rater performed unusually in marking #12 (Shawn and John) rather than only one rater's unusual marking in #20 (Shawn only). It could be speculated that for #12 the unusual marking from these two raters might be due to certain characteristics of this script, which makes it difficult to evaluate its performance. For example, different aspects might have been particularly emphasised by the two raters, which thus affected their marking. The fact that Shawn is the only rater that exhibited such an unexpected marking in #20 makes his awarding this script a higher band score a significant deviation, particular with his being the harshest rater. This is the possible explanation why only #20 rated by Shawn is categorised as the unexpected response by the model but not #12 or #15. The implication for this type of analysis is that we have to consider carefully why certain ratings turn out to be unexpected responses reported by the model. In the case of raters, does the rater in question interpret the rating scale differently with other raters? Does he/she require further retraining if he/she is observed to generate more unexpected responses? In this benchmarking procedure, this irregular marking could probably just be put aside as it is the single case highlighted by the Rasch model among the twenty scripts marked. Moreover, as indicated by the fit statistics, Shawn is still overall labelled as a consistent rater.

6.1.1.3 Benchmarking Meeting & Standard-Setting

Along with the statistics, the face-to-face benchmarking meeting also helped to finalise the CEFR levels assigned to each script and to facilitate standard-setting for the follow-up major rating. A number of interesting points were raised in the discussion. First, the distinction between the two highest levels, C1 and C2, was recognised to be rather difficult for marking. There are only eight instances of C2 among the 100 band scores given (twenty scripts marked by five raters). According to the raters' own accounts at the meeting, one of the two raters who never gave a C2 explained that she was influenced by the pre-benchmarking meeting which had taken place a few months earlier. At that pre-benchmarking meeting, the scripts she awarded a C2 were mostly given a C1 by other raters, which had made her refrain from giving a C2 this time. Another rater remarked that he considered C2 to be a near-native level and was thus conservative in awarding any C2. Additionally, the raters also commented on the application of the CEFR scale. Some doubts were cast on the 'negative' and 'positive' descriptors in the rating scale. For example, the descriptors for the C1 level are overall very positive except for one under the category of 'Range', which claims that 'The flexibility in style and tone is somewhat limited.' This description did not go any further, yet its effects on rating might be worth future investigation. With regard to rater idiosyncrasy, one rater mentioned his attempt to base his judgment upon quantities of various types of errors, but in the end he realised this was an impossible task. He also referred to the scripts he previously marked and tried to make comparison with the script he was marking when in doubt. This practice was disputed by another rater, who advised that raters should always refer to the scale and the standardised scripts rather than making any direct comparison between local performance scripts.

After the benchmarking meeting and statistical analysis, it was then decided that the CEFR levels from A2 to C1 would be the target levels for this study since it was difficult to

find essays at the top level C2 and the bottom level A1.[31] The information provided by the Multi-faceted Rasch analysis as well as descriptive statistics facilitated the selection of scripts representative for each level and suitable to be used as training materials for the major rater training. For the preparation of training materials, a set of sample scripts (seven in total) and two sets of practice scripts (four for each) were selected from benchmarking. The selection process of scripts for rater training materials (standard-setting) can be summarised as below:

- For the sample scripts, two scripts were selected for each level except for A2 with only one sample script, usually the combination of one typical performance (e.g. all the raters gave the same score) and the other less typical (e.g. two raters gave a lower score), so that the raters can have a better understanding of the range in question.
- The priority of arranging the scripts which received the 'better Rasch statistics' and greater rater agreement is samples, practice set 1 and then practice set 2.
- There are no sufficient A2 scripts as during benchmarking only three scripts received A2 ratings. Among the three scripts, #12 registers a potentially problematic script for rating as indicated by the infit statistics and was thus placed in the second practice set. Another two were assigned to the sample set and practice set 1 respectively.

6.1.2 Major Rating

Two experienced raters as well as one of the benchmarking raters participated in the major rating. Different from benchmarking, major rating involves rater training materials from LLC which have been calibrated to CEFR levels (cf. Figure 6-1), and the 1,009 learner scripts in the major rating would be selected to establish the subcorpora representing individual CEFR

---

[31] Although the rater training materials contained only scripts ranging from A2 to C1, the raters were encouraged to note down any scripts that they were uncertain whether the levels fell on A1 or C2. At the end of rating, however, no script received an A1 while there were a number of C2 ratings given.

levels. These essays written by L1 Chinese learners had been preliminarily extracted from the Longman Learners' Corpus to ensure that they are expository or argumentative essays, including some academic essays. The issue of text genres will be further discussed after the statistical analysis of rating results.

6.1.2.1 Procedures

Rater training was conducted through Internet communications because the three raters resided in different areas when this rating task was carried out. As with the benchmarking procedure, each rater received a package of training materials through email, which contained the CEFR rating scale, an instruction sheet, sample essays, and the first practice set. After the raters familiarised themselves with the CEFR rating scale and the sample essays, they started rating the first set of practice essays. The scores given were then discussed with the researcher. Then the raters were given a second practice set and undertook the same procedure again. Given that the raters and the researcher all used to work for the same language testing centre in Taipei where they had received identical rater training, thereby being very familiar with a standard rating procedure, it was not difficult to organise such an electronic rating process via the Internet, and this way of distance training turned out to be fairly effective.

After rater training, all the 1,009 essays that were selected from LLC for rating were disseminated to the raters. These essays were double marked by two of the raters. Any essays which were given two different ratings were then sent to the third rater and marked again. Essays therefore received either two or three ratings – two ratings if the first two raters agreed, and three if the first two raters disagreed. All of the ratings were then aggregated and subjected to statistical analyses in order to investigate inter-rater reliability, to assign a definite CEFR level to each essay, and to decide which essays would be included or discarded for the target subcorpora.

6.1.2.2 Statistics for Major Rating

In this section, various statistical analyses for the LLC rating will be described. As with the benchmarking exercise, descriptive statistics such as standard deviation, means, and inter-rater reliability indexes will firstly be presented. Then the results of applying the Many-faceted Rasch analysis to the rating data will be reported. It will be seen that taking into account the variable of rater severity and consistency, the Rasch analysis can facilitate the process of assigning the most appropriate CEFR level to each essay and selection of essays for investigation.

For the purpose of statistical analysis, the CEFR levels awarded by the raters were again transcribed to numerals: 2 for A2, 3 for B1, 4 for B2, 5 for C1, and 6 for C2. There is no equivalent numeral for A1 because no essay was given an A1 by any of the raters. Two raters Emily and Kat marked all the selected essays, 1,009 in total, from the Longman Learners' Corpus. It should be noted that Kat participated in the benchmarking exercise while Emily did not. A third rater, Penny (who did not involve the benchmarking either), marked 310 essays which the two primary raters disagreed plus 6 agreed ones.[32] As can be seen from Table 6-10, all the essays marked by the three raters received marks from 2 (CEFR A2) to 6 (CEFR C2). The mean scores given by Emily and Kat are very close, 3.85 and 3.84 respectively. As to the standard deviation, the result shows that the dispersion of scores awarded by Emily (0.804) does not vary much with the dispersion of scores given by Kat (0.732). Given that virtually all the essays marked by Penny were the ones disagreed on by Emily and Kat, it is not surprising to see that the descriptive statistics in the scores given by Penny, with a lower mean value (3.72) and a higher standard deviation (0.971), are notably different from those in the two primary raters.

[32] In the beginning, it had been planned to give the third rater more agreed essays for marking for the purpose of comparison, yet later on this idea had to be abandoned owing to the rater's busy schedule.

Table 6-10 Descriptive statistics for major rating

	Mean	Std. Deviation	N	Minimum	Maximum
Emily	3.85	.804	1009	2	6
Kat	3.84	.732	1009	2	6
Penny	3.72	.971	316	2	6

With regard to rater agreement, Spearman's coefficient *rho* and Cronbach's *alpha* index were calculated. The result of Spearman's coefficient *rho* presented in Table 6-11 shows that there was a significant positive correlation between the two primary raters, Emily and Kat (*rho*=0.718, N=1,009, $p<0.0005$, two-tailed). In the context of high-stakes exams, a coefficient higher than 0.8 is generally the aim to be achieved (Davies, et al., 1999, p. 88). However, for high-stakes exams, every writing task is designed with strict specifications and expected to be suitable for candidates with certain proficiency levels. Considering the variability of task types in the Longman Learners' Corpus with written samples mixed with home assignments, academic papers, and test essays, the coefficient 0.718 in the two primary raters is considered acceptable for the purpose of the current study. On the other hand, the coefficients between Penny and the other two raters were 0.551 and 0.594 respectively, which again is due to the fact that Penny mostly marked the essays which received different scores from the two primary raters. The Cronbach's *alpha* index, a measure for internal consistency, is 0.844 between the two primary raters, while the same index is much lower at the level of 0.766 when including the third rater's ratings. The inter-rater reliability indexes reported here are generally lower than the benchmarking (cf. Table 6-3); however, bear in mind that only twenty essays were marked in benchmarking while the major rating included more than 1,000 essays. The drastic difference of amount of ratings involved would be expected to have an impact on the rating performance, thereby revealed in the results.

Table 6-11 Inter-rater reliability for major rating

Spearmans' correlation coefficients		Emily	Kat	Penny
Emily	**Correlation Coefficient**	**1.000**	**.718(**)**	.551(**)
	N	**1009**	**1009**	316
Kat	**Correlation Coefficient**	**.718(**)**	**1.000**	.594(**)
	N	**1009**	**1009**	316
Penny	Correlation Coefficient	.551(**)	.594(**)	1.000
	N	316	316	316
Cronbach's *alpha* index (Emily & Kat): .844				
Cronbach's *alpha* index (three raters): .766				

** Correlation is significant at the 0.05 level (2-tailed).

As has been discussed, neither inter-rater reliability nor descriptive statistics can provide information as to the calibration of ratings considering rater harshness and consistency. Moreover, we have seen that the conventional statistics cannot cope with the third rating when only the disagreed essays were marked. Therefore, the Many-faceted Rasch measurement was introduced again as this measure takes into account both harshness and consistency.

Table 6-12 illustrates a logit scale of candidate ability, rater severity, and the CEFR scale constructed by FACETS. Because the data size is too large to accommodate all the scripts in the table (1,009 scripts), only a selection of results are represented here (including both the top and the bottom of the original table). The far left column (*Measr*) exhibits the spread of candidate ability, rater severity, and the CEFR scale with around 26 logits (+12 to -14). The top of the distribution represents greater candidate ability and greater rater severity. We then see only four candidates with the highest 12 logits, which would be assigned to a solid band score 6 representing the C2 level, while many more candidates clustering around the bottom of the scale with the lowest logit -14 would be a solid band score 2 representing the A2 level. In terms of rater severity, the three raters are virtually positioned at the same point within an acceptable range ±1.0 logit (A. Van Moere, 2006).

Table 6-12 All Facet vertical 'rulers' for major rating

Measr	+candidate								-rater	Scale
	12 +	13071	13078	26396	26397				+	+ (6)
	11 +	11228	13043	13045	13065	26349	26398	457	+	+ ---
	10 +	13077	13079	26395					+	+
	9 +								+	+ 5
	8 +	13038	13039	13041	13042	13044	13047	13048 ...	+	+
	...								+	+ 4
	1 +	11229	13056	13057	13058	13091	13096	13097 ...	* Emily Kat Penny	* *
	-1 +								+	+
	...								+	+
	-8 +	3080	13085	13087	13089	13090	13094	13114 ...	+	+ 3
	...								+	+
	-13 +	1242	1249	22854	22874	22905	29692	29696 ...	+	+ ---
	-14 +	1102	1494	22822	22834	22836	22851	22855 ...	+	+ (2)
Measr	+candidate								-rater	Scale

*Ellipsis indicates that more candidates have been left out owing to the limited space.

Table 6-13 is the rater measurement report in which we will focus on fit statistics only. The column of *Infit MnSq* (Infit Mean Square) shows the most informative fit values. It indicates the relative consistency of ratings. Remember that the rule-of-thumb acceptable range of a fit value falls within 0.75 to 1.3 (McNamara, 1996, p. 173), while Wright and Linacre (1994) proposed an acceptable range of 0.6 to 1.4. For rater consistency, the lower fit figures are generally preferred in the sense that it means the variation between observed and expected values is less than what the model predicts, and hence the rater in question is more consistent. In Table 6-13, Kat is identified with the lowest fit value, 0.85, which suggests that she is the most consistent rater. This is not surprising as Kat was the only rater involved in both benchmarking and major rating. We also see that the fit values of Emily and Penny (1.09 and 1.07 respectively) both fall within the acceptable range. Another rater feature, the degree of severity, was also calculated by FACETS. The fifth column in Table 6-13, *Measure*, presents the extent of rater severity with logits. As we have seen in the FACET vertical rulers, the measures of three raters are very close in terms of rater severity: -0.22 for Emily, -0.08 for Kat, and 0.3 for Penny, all of which fall within the acceptable range of ±1.0 logit. Generally speaking, the fit statistics and the logit measures reported by FACETS provide the evidence

that the markings of LLC essays completed by the three raters are overall not only consistent but also reliable.

Table 6-13 Rater Measurement Report

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	N rater
3888	1009	3.9	3.89	-.22	.10	1.09	1.4	.73	-1.6	.99	.91	.48	61.0	74.8	1 Emily
3874	1009	3.9	3.87	-.08	.10	.85	-2.4	.60	-2.5	1.14	.91	.48	63.9	74.9	2 Kat
1177	316	3.7	3.82	.30	.13	1.07	1.2	2.28	9.0	.32	.83	.38	44.6	58.4	3 Penny

In addition to providing information about the raters, FACETS also generates a candidate measurement report, which greatly facilitates the determination of a final CEFR level for each LLC essay. Table 6-14 shows a selection of the candidate measurement report produced by FACETS. The values in two columns, *Infit MnSq (*Infit Mean Square) and *Fair-M Avrage* (Fair-M Average), provide valuable information as to the fairness of rating received in each piece of learner writing.

Table 6-14 Candidate Measurement Report

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Num	candidate
4	2	2.0	2.13	(-15.11	1.98)	Minimum					.00	.00	637	637
7	3	2.3	2.33	-13.93	1.23)	.87	-.1	.84	-.2	1.57	.84	.03	1102	1102
6	2	3.0	3.00	-7.72	7.75	.00	2.2	.00	2.2	1.18	.00	.00	23014	23014
9	3	3.0	3.00	-7.86	6.26	9.00	3.6	9.00	3.6	-13.0	.77	.01	7784	7784
11	3	3.7	3.67	-1.56	1.22	3.71	3.9	3.87	3.9	-13.1	.70	.03	7721	7721
14	3	4.7	4.67	5.47	1.21	1.05	.2	1.05	.2	.93	-.21	.03	24439	24439
14	3	4.7	4.67	5.47	1.21	4.33	4.1	4.18	3.9	-9.42	-.94	.03	24702	24702
16	3	5.3	5.33	10.01	1.21)	1.19	.5	1.25	.6	.54	-.90	.03	13077	13077

The first step is to look at the fit statistics in the column of Infit Mean Square. Remember that a universally acceptable fit range is 0.75 to 1.3, which indicates the reasonable extent of variation expected by the Rasch model. If the fit statistic is too high, it means the variation of ratings has gone beyond what is expected by the model (*misfit*). If the

fit statistic is too low, then it shows lack of variation (*overfit*). Any rated LLC essay with the Infit Mean Square value falling out of this range was therefore examined and checked against the original CEFR levels awarded by the raters. This examination shows that an Infit value greater than 1.3 refers to those essays that were disagreed on significantly (statistically speaking) by the raters. For example, as shown in Table 6-14, Candidate No.7784 (for candidate number, see the far right column) with an extremely high Infit value 9 was awarded three different band scores by the three raters respectively: A2, B1, and B2. In the context of real-life language tests, for an unusual rating like this, a general practice would involve trying to get a fair mark for this piece of candidate performance. For the present study, however, the 'misfitting' essays with over-high Infit values might be a concern for the follow-up linguistic investigation on the grounds that some underlying aspects of these essays caused the experienced raters to mark them so differently to the extent that they are regarded as 'misfits' by the model. Candidate No. 7721 is another good illustration of this as it received two non-adjacent levels from the three raters, i.e. C1, B1, and B1, thus with a high Infit value 3.71. On the other hand, essays with an Infit value lower than 0.75, the 'overfitting' ones, were found to be given the same band scores by the two primary raters. The essays with an Infit value within 0.75 to 1.3 are generally the ones which were agreed on by two raters and disagreed by the third rater by one band score. For example, essay No.1102 with the Infit value of 0.87 received the band scores B1, A2, and A2 from the three raters. It was thus decided to keep any essay with an Infit value below 1.3 for the time being and exclude those with a Infit value higher than 1.3, which either suggests erratic rating behaviour or atypical performance for the particular level, from the constructed corpus. Overall speaking, only 34 out of 1,009 essays were defined as 'misfits'. In other words, the majority of rated LLC essays (96.6%) are considered to be within reasonable scoring variation.

After filtering out the essays which failed to meet the criterion of acceptable fit values,

the next step is to examine the candidates' ability estimates, which are indicated by the fourth column in Table 6-14, *Fair-M Avrage*. If the value of ability estimate is not an integer, then it would be rounded up or down subject to whether the decimal fraction is higher or lower than 0.5 (there were no instances of ability estimates receiving a mark ending in .5 like 2.5 or 3.5). As can be seen from the examples in the table, the difference between raw scores (*Obsvd Average*) and Rasch-adjusted scores (*Fair-M Avrage*) are actually minimal.

The rated essays were examined and assigned to a final CEFR level on the basis of the above described fit statistics and ability estimates generated by FACETS. Unfortunately, only the CEFR C1 and B2 levels encompass a substantial number of essays that can be considered marginally sufficient to construct a solid corpus (cf. individual subcorpora of 100,000-200,000 words by learners of different L1s in the International Corpus of Learner English (ICLE)), especially for C1 level writing. As Table 6-15 indicates, the C1 subcorpus contains 157 pieces of learner writing with a total word count of 87,828 while the B2 subcorpus consists of 469 pieces of learner writing with 158,489 words in total. The other three levels are comparably rather small, particularly at the top and bottom levels C1 and A2. Moreover, text length also appears to be proficiency-dependent, which reflects the fact that proficient learners tend to produce (or the set essay tasks require more proficient learners to produce) longer essays. In an ideal comparative study which intends to investigate learner language development, the learner corpora representing proficiency levels would be constructed in such a way that corpus size is as similar as possible and so should be the number of texts.[33] The result revealed here, nonetheless, suggests that this is highly unlikely to be achievable.

---

[33] Oakey (2009), nevertheless, pointed out that text length does not affect retrieval of lexical bundles. The interaction between variables of corpora and lexical bundles will be further discussed in Chapter 7.

Table 6-15 Corpus size and number of texts built on the basis of Rasch-calibrated ratings[34]

Rasch-calibrated CEFR Level	**Corpus size** (word count)	**No. of texts**	**Average text length**
C2	3,922	7	560
C1	87,828	157	559
B2	158,489	469	337
B1	44,941	236	190
A2	1,275	11	116

As has been discussed throughout this section, rating is a complex process which requires a great deal of cautious planning and decision-making. Subjectivity is unavoidable in assessing linguistic performance. Rater training is indispensable in that it can calibrate raters' judgment and hopefully eliminate inconsistency to the minimum. Discussion with raters is also useful in understanding their rating behaviour as raters might have different concerns about any of the variables involved during the rating process, e.g. essays or the rating scale. The Multi-faceted Rasch analysis helps the benchmarking to be conducted more efficiently by not only identifying but also quantifying the extent of disagreement of ratings. It is also conducive to determining the degree of rater severity and consistency, two crucial characteristics for determining a good rater. Furthermore, the Rasch analysis can spot any problematic ratings that may result in controversy, which are otherwise very difficult to detect via human eyes, especially with a large amount of data. Although the Rasch analysis is powerful in analysing marking behaviour, however, further explanation about unexpected ratings still requires the researcher to look into the data and seek the most appropriate interpretation.

[34] A preliminary comparison between the final CEFR levels and the original Longman annotation regarding proficiency levels is indicative of drastic difference between these two sets of proficiency information.

6.2 Selection of Data

After major rating was complete, the rated essays were further selected for Modular Study 2, which requires learner essays representative for different CEFR proficiency levels. There were two stages for this selection process, one based on the Rasch analysis of rating results and the other on the examination of text/task types. As discussed in the last section, the Rasch analysis computed by FACETS has shown that the rating of the LLC essays, as a whole, is satisfactory in terms of rater severity and consistency, and the fit statistics for the rated essays also facilitates the determination of data for each proficiency level to a very large extent.

On the other hand, during the process of rating, the issue of text/task type variation was noticed although the LLC essays had been preliminarily examined and selected before rating. After the completion of rating, it was decided to scrutinise the tasks of those selected essays in the subcorpora again with a much more thorough approach. This time the documentation of the writing tasks was taken into account, and only those texts that were argumentative or expository in nature were kept in the corpus.[35] The aim of this second scrutiny was to ensure the comparability of task forms between essays. To be more specific, argumentative essays present the writers' arguments as well as facts, intending to persuade the readers with regard to a particular issue. Most of the selected learner essays are of this type. In comparison, expository essays refer to those which give information and facts rather than the writers' opinions. Some essays, however, appear to span across both text types and

---

[35] When the essays were chosen the first time for rating, very often only the titles or essay questions that learners responded to were examined so as to determine whether they were the target task types as there was a huge amount of data to undertake such an examination (all the L1 Chinese essays of L2 English from LLC). However, it has been found that essay titles or questions are not a very reliable indicator for task types. The second scrutiny described here, therefore, involves reading through each text as opposed to a quick scan of the essay titles or questions only.

are thus labelled as 'Argumentative/Expository'. Academic essays identified as university assignments were also included because they correspond to the profiles of expository or argumentative essays: they present either information or an argument and very often a mixture of both. Another practical reason to include academic essays is simply to keep the corpus size as large as possible. The specialised terminology used in academic writing, as described in Section 3.6.1 about context-independency, would be removed from the extracted lexical bundles and therefore would not affect the investigation. Some examples of the titles or questions the learner essays responded to in each text type will be illustrated in Section 6.3.4.

The essays which appeared to have an expository or argumentative title but in terms of content employed a personal tone instead were discarded. For example, some learners responded to a supposedly argumentative essay title such as 'You Are What You Read and Experience' but produced a piece of narrative essay by using scenarios of their daily life as examples throughout the text. Such deviating forms, all found at the lower B2 level, were considered to be incomparable with the expository or argumentative essays targeted in this study as they solicited different types of language. Moreover, expository and argumentative essays could be more comparable with the investigation of academic writing in Modular Study 1 although it still has to be borne in mind that they do not represent the same genres. As discussed earlier in Chapter 4, the mixture of text types investigated in Modular Study 2 is therefore termed as 'EAP-like writing' as they are very much the same as the text types that most EAP programmes are concerned with.

The last concern is corpus size. Considering the amount of data that would be suitable for retrieval of lexical bundles for comparisons across levels, only the subcorpora B2 and C1 would be investigated for Modular Study 2 as they are the only two groups of writing which have a substantial number of texts appropriate for corpus research. However, given

that the corpus size of B2 and C2 is still quite different, task type was used as a criterion to not only further downsize the B2 subcorpus but also ensure that the distribution of text types in the B2 and C1 subcorpora is as comparable as possible. The finalised constituents of these two rated corpora can be seen in the table and figure below.

Table 6-16 Corpus size and number of texts of the original rated data and the selected rated data

	Original rated data		Selected rated data	
CEFR Level	**Corpus size** (word count)	**No. of texts**	**Corpus size** (word count)	**No. of texts**
C1	87,828	157	87,828	157
B2	158,489	469	87,970	239

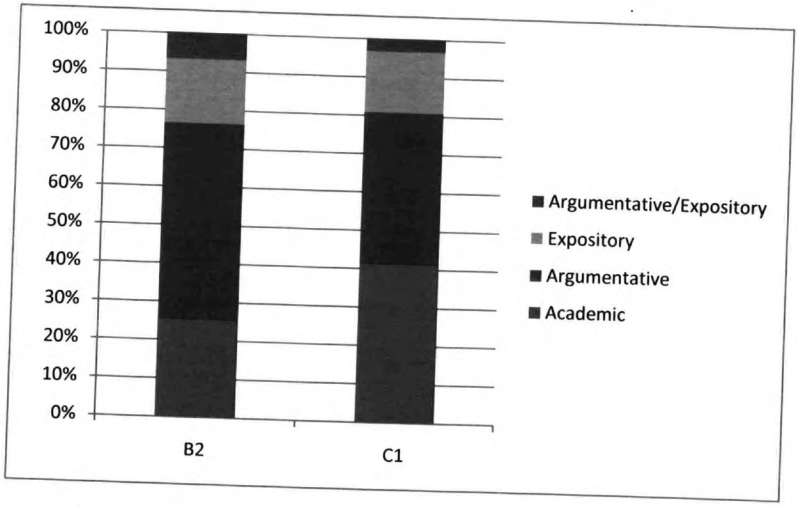


Figure 6-2 Text types in the finalised CEFR-B2 corpus and CEFR-C1 corpus

To sum up, these two finalised CEFR-standardised subcorpora retrieved from the Longman Learners' Corpus have the following characteristics:

1) they were all written by L1 Chinese learners of L2 English;

2) they were rated and agreed upon by at least two raters;

3) they consist of argumentative, expository and academic essays; and

4) they are of comparable corpus size in spite of different numbers of texts encompassed.

6.3 Linguistic Profile

In this section, further information with regard to the two constructed subcorpora B2 and C1 will be addressed. Different from Modular Study 1, the learner data extracted from the Longman Learners' Corpus is much less controlled and thus a post-hoc analysis of learner backgrounds and linguistic features might help us better understand the texts chosen for these two groups of writing of different levels.

6.3.1 Learner Backgrounds

In the Longman Learners' Corpus, some contextual information, such as the learners' mother tongues or task types, is coded in the header of each text. Although the level of each written sample was also tagged, the process used by the LLC compilers in order to determine learners' proficiency was not documented. Despite repeated attempts, it has not been possible to get in touch with the compilers of the corpus or anyone else who can answer questions regarding proficiency rating. This is also one of the reasons why I decided to adopt a robust rating procedure to define learner proficiency in the current study.

Some information about learners[36], however, is retrievable from the documented annotation. After a review of all the contextual information, here only learners' L1 and its variety will be presented as other important variables such as text types have been considered when constructing the CEFR subcorpora. For the issue of complexity of L1 Chinese variety and learning of L2 English, please see Section 2.1.1. Since all these learner essays have been

[36] Learners' age or education backgrounds are not documented in the Longman Learners' Corpus.

selected through rating, the origins of areas where the L1 Chinese learners came from does not seem as important. Yet as can be seen from Figure 6-3, the learner essays contributed from Hong Kong are the majority in each of the learner subcorpora, which explains why a large number of context-dependent bundles relating to learners' backgrounds *Hong Kong* were retrieved (cf. Section 3.6.1). In addition, we can also see that the learners from the three L1 Chinese varieties show a similar distribution in CEFR-B2 and in CEFR-C1 subcorpora.[37]

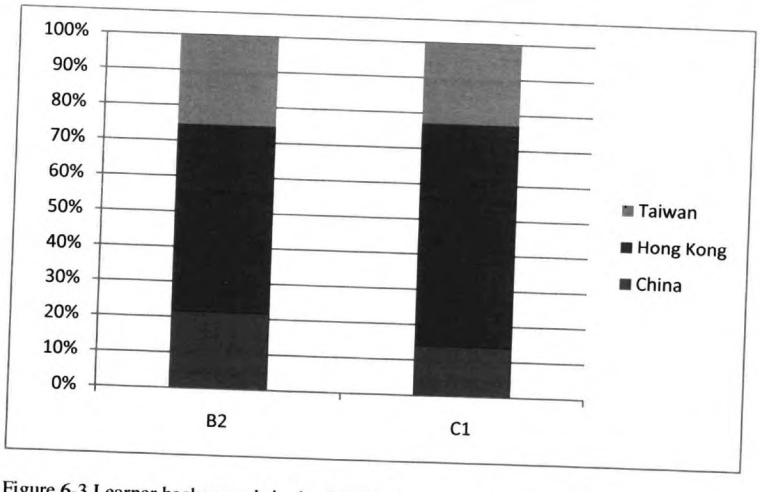


Figure 6-3 Learner backgrounds in the CEFR-B2 corpus and CEFR-C1 corpus

6.3.2 Type/token Ratio

As addressed in Section 5.2.1, the standardised type/token ratio (STTR) provided by the *WordSmith* wordlist, which computes the mean type/token ratio (TTR) by breaking up a

[37] Since each script in the Longman Learners' Corpus was coded with a reference number and some background information about the learner who contributed the script, unfortunately, it is impossible to know whether the learners contributed more than one piece of writing. As the data used for this Modular Study only forms a limited range of scripts from a very large amount of data available with multiple selection processes, it can only be assumed that the chance of choosing more than one script from the same learner has probably been reduced.

corpus into 1,000-word chunks, can be used to reflect 'lexical density' and thus to compare different sets of data. It is generally assumed that the higher the ratio, the more varied the lexis used in the text. In Table 6-17, nonetheless, the STTR value at the lower level CEFR-B2 turns out to be slightly higher than that at the C1 level. There might be various explanations for this. As Table 6-16 shows, the CEFR-B2 subcorpus contains nearly 80 more essays than the C1 subcorpus (239 vs. 157), and hence it probably has a wider range of task types and essay titles, which could lead to elicitation of a much wider range of lexis. Indeed, just as with any other measures of lexical variation (e.g. the ratio of lexical words to overall words), STTR does not seem to be exempt from variables concerning corpus constituents such as numbers of texts and hence task types here. At least, it is clear that STTR does not appear to be an effective measure in distinguishing proficiencies in corpus-based studies of this kind if it is sensitive to these text-related variables.

Table 6-17 STTR in Modular Study 2

	CEFR-B2	***CEFR-C1***
STTR (per 1,000 words)	7.5	7.25

Interestingly, in comparison with the STTR values in the three corpora in Modular Study 1 (see Table 6-18), a sharp contrast can be seen between the two types of genres and/or proficiency: one is academic writing with longer text length and average STTR as high as 39 and the other is EAP-like learner writing with generally shorter text length and STTR ranging from only 7.25 to 7.5.

Table 6-18 Reproduction of Table 5-7 (STTR in Modular Study 1)

	BAWE-CH	***BAWE-EN***	***FLOB-J***
STTR (per 1,000 words)	39.3	39.16	39.64

6.3.3 Words with Different Lengths

As in the previous chapter, some information about word length is provided in order to help us better understand the language used in the two constructed CEFR subcorpora. In Table 6-19, we can see that CEFR-B2 learners tended to use words with shorter lengths (from 1-letter to 6-letter words) while from 7-letter words on, CEFR-C1 learner writing generally had more occurrences of words than CEFR-B2 writing. Although this tendency is not supported by the three corpora in Modular Study 1, it seems an interesting pattern here which might be worthy of future research.

Table 6-19 Numbers of words with different lengths in Modular Study 2 (occurrences)

N	CEFR-B2	CEFR-C1
1-letter words	3,170	2,441
2-letter words	15,960	15,681
3-letter words	18,366	16,364
4-letter words	15,746	13,088
5-letter words	9,791	9,195
6-letter words	7,617	7,074
7-letter words	7,093	7,460
8-letter words	4,557	6,165
9-letter words	3,516	3,987
10-letter words	2,410	2,728
11-letter words	1,271	1,940
12-letter words	601	898
13-letter words	358	557
14-letter words	123	147
15-letter words	29	61
16-letter words	14	12
17-letter words	4	16
18-letter words	1	10
19-letter words	0	0
20-letter words	0	1

6.3.4 Topics of the Written Samples

This section exemplifies the topics or questions from the learner essays assigned to two CEFR subcorpora B2 and C1. Two points have to be noted. Firstly, as emphasised earlier, the text type was not determined by the essay titles or questions alone but by an overall examination of each piece of text. Secondly, there is no clear distinction of the topics or questions between the B2 and C1 subcorpora. As can be seen from Table 6-20, many of the topics or questions are similar in these two groups of learner writing. In fact, in order to cover a wider range of tasks, if the topics or questions are shared by these two subcorpora, often only one set of them are present in either CEFR-B2 or C1 in this table. Therefore, it could be claimed with some confidence that such a careful post-hoc examination and selection of task/text types should be able to minimise the impacts from using data with less control of various variables such as data from the Longman Learners' Corpus.

Table 6-20 Some examples of essay topics in Modular Study 2

Text type	CEFR-B2	CEFR-C1
Argumentative	• Greatest Problem Facing the World • Role of Women • Every day life is no longer possible without computer	• Computer • To what extent do the benefits of study abroad justify the difficulties involved? • Abortion
Expository	• Place of Historical Interest • Geography of Hong Kong • Changes In British Advertising in the Last 30 Years	• Communication within the Chinese Class • History of Hong Kong • Newspaper
Argumentative/ Expository	• Past and future of Hong Kong • Describe something important that happened in your country's history: how did it happen, and why was it so important? • Management And Rules of Chai-Shun National Park	• Good and bad features of life in your home town • Common agricultural policy • Education in Hong Kong
Academic	• Barristers and solicitors • Credit Behaviour of Peasants in Rural China • First and Second Language Acquisition	• Barristers and solicitors • Conventional Marxism • Mystical Numbers and Manchu Traditional Music • First and Second Language Acquisition

6.4 Structural Analysis

Just as in Modular Study 1, two kinds of structural analysis, type and token distribution, were carried out. The lexical bundles extracted from the two CEFR subcorpora were categorised into NP-based, PP-based, and VP-based bundles as described in Section 4.2. Yet five of these bundles allocated to the structural category of 'Others' had to be excluded in the quantitative analysis because they do not constitute a sufficient amount of data in individual subcorpora for a chi-square test. These bundles are *him or her to* found in CEFR-B2 and *as far as the*, *as well as the*, *how to deal with*, and *necessary for us to* found in CEFR-C1. Altogether they account for only 4.6% of the total 108 bundle types in Modular Study 2.

6.4.1 Type Distribution

As can be seen from Table 6-21 and Figure 6-4, VP-based bundles are the majority in both CEFR-B2 and CEFR-C1 learner writing, but the proportion in CEFR-B2 (65.7%) is higher than that in CEFR-C1 (51.5%). With regard to the other two categories which do not contain any verb elements, the more proficient learner writing in CEFR-C1 has a higher proportion of PP-based bundles (30.3%) whereas the less proficient writing in CEFR-B2 has a higher proportion of NP-based bundles (21.4%).

Table 6-21 Structural distribution in Modular Study 2 (types)

χ^2=4.241, df=2, p=0.12			Structure			Total
			NP-based	PP-based	VP-based	
Corpus	CEFR-B2	Count	15	9	46	70
		% within Corpus	21.4%	12.9%	65.7%	100.0%
	CEFR-C1	Count	6	10	17	33
		% within Corpus	18.2%	30.3%	51.5%	100.0%

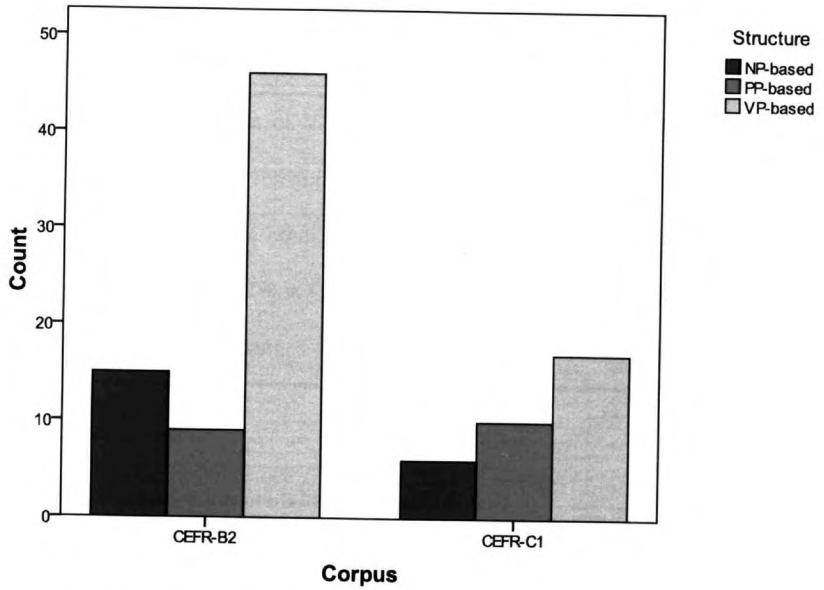


Figure 6-4 Structural distribution of lexical bundles in Modular Study 2 (types)

However, the chi-square test shows that there is no significant difference in terms of structural distribution of bundle types between CEFR-B2 and CEFR-C1 writing at the 0.05 level (χ^2=4.241, df=2, p=0.12)[38], although there appears to be noticeable difference in the proportion of the three structural categories. We shall see whether the token distribution will mirror the same pattern of structural distribution between CEFR-B2 and CEFR-C1 after frequency is taken into account in the next section.

---

[38] Again, the Bonferroni Correction is not considered in this chapter. The purpose of such correction is to guard against spurious positives resulting from multiple comparisons. However, as we shall see later, the p values from structural and functional distributions in this chapter are either very high (i.e. 0.12, 0.781, and 0.07) or very low (p<0.0005), and thus the results concerning significance are fairly clear without the need to further complicate the analysis by lowering the set significance level .

6.4.2 Token Distribution

As can be seen in Table 6-22 and Figure 6-5, the structural distribution of bundle tokens between CEFR-B2 and CEFR-C1 is similar to the type distribution, only with some slight variation. The proportion of VP-based bundles drops to 59.7% in CEFR-B2 and 45.9% in CEFR-C1, yet they both remain the category that holds the most occurrences in either CEFR-B2 or CEFR-C1. On the other hand, PP-based bundles slightly increase in both groups of learner writing, with 18.2% in CEFR-B2 and 36.9% in CEFR-C1 while the proportion of NP-based bundles rarely changes within the same corpus (22.1% in CEFR-B2 and 17.1% in CEFR-C1).

Table 6-22 Structural distribution in Modular Study 2 (tokens)

χ^2=27.093, df=2, p<0.0005			Structure			Total
			NP-based	PP-based	VP-based	
Corpus	CEFR-B2	Count	90	74	243	407
		% within Corpus	22.1%	18.2%	59.7%	100.0%
	CEFR-C1	Count	38	82	102	222
		% within Corpus	17.1%	36.9%	45.9%	100.0%

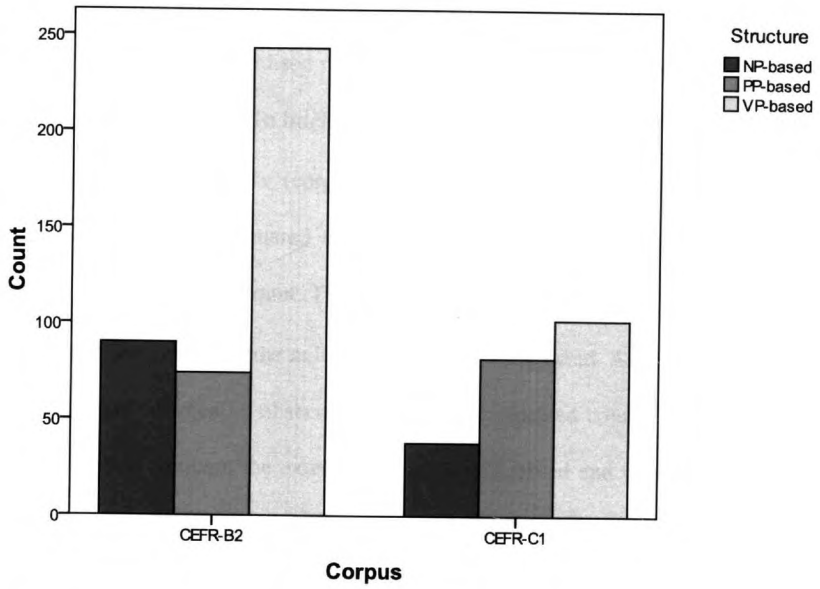


Figure 6-5 Structural distribution in Modular Study 2 (tokens)

The chi-square test indicates that there is significant difference in terms of structural distribution of bundle tokens between CEFR-B2 and CEFR-C1 writing at the 0.05 level (χ^2=27.093, df=2, p<0.0005). The significant difference disclosed in the token distribution, a contrastive manner of type distribution, suggests that the difference of use of lexical bundles in two groups of learner writing becomes more pronounced when the variable of frequency is taken into account (because χ^2 partly hinges on sample size). The similarity between CEFR-B2 and CEFR-C1 writing is that VP-based bundles continue to make up the majority of bundles regardless of type or token distribution. Meanwhile, the different degree of reliance on PP-based bundles is also worth more attention, which will be further discussed in the statistical tests in the next section.

6.4.3 The Chi-square Test and Standardised Residuals

This part of analysis aims to find out which structural categories serve to better distinguish

groups of written samples in terms of writing competency. The procedure of calculating a chi-square test and the standardised residuals has been thoroughly discussed in Modular Study 1, Sections 5.3.3 and 5.4.3. To briefly recap, the chi-square statistic (χ^2) equals to the sum over all elements in the matrix (corpus by structural category in this case), and each of the elements refers to the squared difference between the observed and the expected counts divided by the expected count. The higher the chi-square statistic χ^2, the lower the critical p value and the more likely the null hypothesis will be rejected. As for residuals, they represent the difference between an observed count and an expected count. Standardised residuals (R) can be used to compare the extent to which the observed and expected counts differ across contingency tables. With an absolute value of R for a given cell greater than 1.96, it means that the cell in question makes a major contribution to rejecting the null hypothesis.

In this section, the chi-square test was again executed with the two subcorpora, CEFR-B2 and CEFR-C1 learner writing. The focus would be cast on which elements/cells in the two (corpus) by three (structural category) table contribute more to the chi-square statistics χ^2 among the two groups of writing so that they may function as a good indicator for distinguishing writing competency. For this purpose, although the test shows that there is no significant difference in the structural distribution of bundles types between CEFR-B2 and CEFR-C1 writing, chi-square residuals were still calculated so as to reveal which cells in the chi-square contingency table differ more than the others. The results of standardised residuals R in Table 6-23 indicates that the two cells representing PP-based bundles have the two values with the greatest magnitude, -1.1 and 1.6, which suggest that CEFR-B2 learners used fewer PP-based bundles than expected while CEFR-C1 learner writing contains more PP-based bundles than expected. Yet it should to be noted that these values do not reach the threshold 1.96 to suggest a substantial contribution.

Table 6-23 Chi-square standardised residuals for structural distribution (types) in Modular Study 2

χ^2=4.241, df=2, p=0.12		NP-based	PP-based	VP-based
CEFR-B2	Observed Count (O_i)	15	9	46
	Expected Count (E_i)	14.3	12.9	42.8
	R	0.2	-1.1	0.5
CEFR-C1	Observed Count (O_i)	6	10	17
	Expected Count (E_i)	6.7	6.1	20.2
	R	-0.3	1.6	-0.7

As to token distribution, the result shown in Table 6-24 reveals that the values with the greatest magnitude, 3.6 and -2.7, were again found in the two cells representing PP-based bundles, which both make a substantial contribution to the significant difference at the 0.05 level. Once more, the negative *R* value suggests that CEFR-B2 learner writing has fewer PP-based bundles than expected whereas the positive *R* value CEFR-C1 suggests otherwise.

Table 6-24 Chi-square standardised residuals for structural distribution (tokens) in Modular Study 2

χ^2=27.093, df=2, p<0.0005		NP-based	PP-based	VP-based
CEFR-B2	Observed Count (O_i)	90	74	243
	Expected Count (E_i)	82.8	100.9	223.2
	R	0.8	**-2.7**	1.3
CEFR-C1	Observed Count (O_i)	38	82	102
	Expected Count (E_i)	45.2	55.1	121.8
	R	-1.1	**3.6**	-1.8

Overall speaking, the use of PP-based bundles appears to contribute the most to the difference between less proficient CEFR-B2 writing and more proficient CEFR-C1 writing in the chi-square test in both type and token distribution. The category of VP-based bundles also makes some contribution to the difference in both type and token distribution, although the magnitude does not reach the required threshold of 1.96. In the next section, we shall move on to see which PP-based and VP-based bundles in these two CEFR subcorpora lead to the

quantitative difference revealed here.

6.4.4 Use of Lexical Bundles in Structural Categories

As addressed in Section 4.2, each of the three major structural categories, NP-based, PP-based, and VP-based bundles, is further divided into a number of subcategories according to the grammatical forms subsumed. In this section, I look into the qualitative difference between CEFR-B2 and CEFR-C1 writing from the perspective of structural categorisation. More often than not, these subcategories include fewer than five instances in individual subcorpora, and hence they are not suitable for the chi-square test for quantitative analysis.

6.4.4.1 NP-based and PP-based Bundles

NP-based bundles are composed of nominal phrase fragments with *of* (NP+*of*) and any other nominal phrase fragments without *of* (NPf); likewise, PP-based bundles comprise prepositional phrase fragments with *of* (PP+*of*) and any other prepositional phrase fragments without *of* (PPf). In the NPf subcategory, it was found that four out of six bundles in CEFR-B2 learner writing subsume a quantifier *a lot of*: *a lot of people*, *a lot of problem(s)*, *a lot of time*, and *and a lot of*. The concordance lines indicate that these bundles were mostly retrieved from different texts, which means the use of *a lot of* is widely spread among CEFR-B2 writing as opposed to some idiosyncratic style only found in certain learners. The salient overuse of the quantifier *a lot of* might be due to learners' tendency of overstatement in writing. Lorenz (1998) has pointed out the characteristic of undue adjective intensification in advanced learner writing, concluding that learners might attempt to stress their argument and impress the reader by means of excessive use of intensifying adjectives. Although the quantifier *a lot of* is not within the domain of intensifying adjectives that Lorenz investigated, it is believed that such tendency of overstatement might be demonstrated in other lexico-grammatical forms in learner writing. This assumption can be evidenced in the rest of this

chapter and will be further discussed in Chapter 7. In addition, this inordinate use of *a lot of* was not observed in CEFR-C1 writing, which suggests that this stylistic infelicity seems to improve as learner proficiency progresses.

Within all the NPf bundles with *a lot of*, one concern with regard to the methodological issue was brought forward by the highly frequent expression *a lot of problem(s)*, which has a bracket with a plural suffix *-s* added to the end of *problem* because this expression includes six occurrences of the erroneous form **a lot of problem*. As the only lexical bundle derived from an error found in the bundle repertoire, it highlights a potential caveat of applying such a quantitative approach to the interlanguage: any spelling or grammatical mistakes hidden in the learner language might have impacted the automatic retrieval of lexical bundles and thus affected the outcome. Say one four-word combination occurring four times in three texts is supposed to be included in the generated list, but one spelling mistake in one of the occurrences of this four-word combination would impede the computation to count it in because the computer only retrieves the word sequences with exactly the identical forms. It is very likely that certain word sequences might have been under-represented because of learner-specific deviant forms. An ideal learner corpus, therefore, should be one which has been error-tagged with both the original erroneous form and the corrected form in a way that does not hinder the automatic retrieval of recurrent word sequences. As the Longman Learners' Corpus is not error-tagged and error-tagging is not within the scope of the current study, such a possible flaw in the data has to be acknowledged, and it will also be one direction for future research. For the current study, it has been found that certain bundles are particularly more error-prone, and we shall see more examples in the rest of this chapter.

If we turn to fixed frames (or termed as 'phrase-frame' by Stubbs, 2007a), remember the two frames derived from NP-based and PP-based bundles that occur very frequently in

academic writing: '*the* + Noun + *of the/a*' and '*in the* + Noun + *of*', which were discussed in Modular Study 1 (cf. Table 5-15 and Table 5-16). For the first frame, '*the* + Noun + *of the/a*', a number of lexical bundles in CEFR-B2 and CEFR-C1 were found to fit into it, and these two groups of learner writers as a whole almost use the same range of nouns for this frame regardless of their proficiency levels (see Table 6-25). The second frame, '*in the* + Noun + *of*', was absent in these two learner subcorpora. Compared with the findings in Modular Study 1, these two productive 'fixed frames' in academic prose turn out to be not productive at all in general learner essays, at least not in the samples from the Longman Learners' Corpus.

Table 6-25 The frame '*the* + Noun + *of the/a*' used in Modular Study 2

the* + Noun + *of the/a		**Total**	
		type	**token**
CEFR-B2	**end** (4), **quality** (4), **rest** (4), result (9)	4	21
CEFR-C1	**end** (6), **quality** (6), **rest** (6)	3	18

* The lexical bundles appearing in both CEFR-B2 and CEFR-C1 are indicated in bold print.

In terms of PP-based bundles, there are six lexical bundles in the subcategory of the prepositional phrase fragments without *-of* (PPf) in CEFR-B2 and in CEFR-C1 respectively. Five of the PPf bundles in these two groups are identical: *as a matter of (fact)*, *for a long time*, *all over the world*, *at the same time*, and *on the other hand*. More interestingly, three of these PPf bundles also appear in the three corpora of academic writing in Modular Study 1. Table 6-26 shows all the PPf bundles in this modular study and the corresponding frequency shared by the five groups of writing. For the two bundles shared by all the five groups, *at the same time* and *on the other hand*, it is clear that all the L2 learner groups (CEFR-B2, CEFR-C1, and BAWE-CH) used them far more frequently than the two native groups (BAWE-EN and FLOB-J). Combined with the frequency of the other bundles in Table 6-26, it can be assumed that learners appear to favour a number of formulaic expressions such as *as a matter of fact*,

all over the world, *at the same time* and *on the other hand*, and they used them more often than native speakers did. The tendency of overstatement in learner writing was also observed again in some adverbial expressions such as *for a long time* and *all over the world*, which has not appeared in the native writing investigated. This suggests that learners' tendency of overstatement may be embodied not only in intensifying adjectives as suggested by Lorenz (1998) but also in certain adverbial phrases with the intensifiers that modify extent or quantity (e.g. *a lot of*, *long*, *all*). From such kind of observation and comparison with Modular Study 1, some interesting patterns shared by different groups of learner writing have emerged despite genre difference. We shall see an overall comparison with distinct features across writing development in Modular Studies 1 and 2 in the next chapter.

Table 6-26 Lexical bundles and the frequency in the subcategory of PPf bundles in Modular Studies 1 & 2

Corpus / (word count) / Bundle / Freq	Modular Study 2		Modular Study 1		
	CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
as a matter of (fact)	4	5	--	--	--
for a long time	11	6	--	--	--
in the following paragraphs	4	--	--	--	--
in such a way (that)	--	5	--	--	--
all over the world	5	5	6	--	--
at the same time	17	14	24	5	10
on the other hand	18	28	36	4	19

6.4.4.2 VP-based Bundles

Four subcategories under VP-based bundles contain noticeably more bundles in CEFR-B2 writing than those in CEFR-C1. These subcategories are 'adverbial clause', 'S+V', 'V*to*' and 'VPf' (see Table 6-27).

Table 6-27 Numbers of bundles in VP-based subcategories

Corpus		VP-based bundles								Total
		adverbial clause	Copula BE	it	PasPP	S+V	to	Vto	VPf	
CEFR-B2	type	4	4	7	1	17	1	7	3	44
	token	16	30	35	5	87	4	48	18	243
CEFR-C1	type	0	2	6	1	5	2	1	0	17
	token	0	19	36	4	30	9	4	0	102

The four bundles under the 'adverbial clause' subcategory in CEFR-B2 writing all begin with a subordinator: *as I have mentioned*, *as we all know*, *because they are not*, and *if there is a*. The use of the first person pronouns *I* and *we* in the adverbial-clausal bundles are also quite prevalent in the 'S+V' subcategory in CEFR-B2 writing. Five out of the 17 'S+V' bundles contain the first person singular pronoun *I* or the plural pronoun *we*: *I am going to*, *I think it is*, *I think that this*, *I would like to*, and *we can see the*. Three out of the five 'S+V' bundles in CEFR-C1 writing have first person pronouns as the subject, too: *I would like to*, *we can say that*, and *we can see that*. These bundles were compared with those with the identical identification/focus function in academic writing (*we can see that* and *it can be seen that*). In the examples below, we see a shift of perspective from being personally involved towards being more impersonal via the change from the use of the writer's perspective alone *I* to the allied relationship with the reader *we* towards the impersonal pronoun *it*. The implication is that as the competency of the writers develops, the tone tends to become more depersonalised. However, the genre difference could also contribute to the de-emphasis of the writer as impersonalisation is viewed by many textbooks or style guides as one distinctive

convention in academic discourse (Hyland, 2002, p. 1095).

- ***I think that this*** is the key reason why the government investment is so inefficient, not the critical reason of the high share of investment. (CEFR-B2)
- Most people think it is the responsibilities of women to clean their home and take care of their children. But ***I think it is*** completely wrong. (CEFR-B2)
- Moreover, the animal such as apes use both the visual and vocal languages to communicate. From this point, ***we can see that*** human also uses both the visual and vocal languages to communicate. (CEFR-C1)
- If we take history into an account in the effect of the economic development, ***we can see that*** the cold war gave East Asia a chance to develop. (BAWE-CH)
- Having weighed up the arguments, ***it can be seen that*** there is an element of truth in this statement. (BAWE-EN)
- ***It can be seen that*** both the main effects are significant; sex at the 0.01 level and faculty at the 0.05 level. (FLOB-J)

The rare use of passive voice in the VP-based bundles is also noteworthy. There are two different bundles being assigned under the subcategory of 'passive verb + prepositional phrases' (PassPP) in these two CEFR subcorpora (*is based on the*, *can be divided into*). After a close examination of the VP-based bundles, another two bundles with the passive voice are located from other subcategories (*are not allowed to*, *it is believed that*), yet altogether there are only four bundles with a passive verb in the whole repertoire (see Table 6-28). As discussed in Modular Study 1, L2 students did not use the passive-verb forms in lexical bundles as frequently as the two native groups (cf. Section 5.3.4.2 and Table 5-21). The use of passive verbs, in fact, can be associated with the impersonalised voice, which has just been

discussed earlier. Overall speaking, impersonalisation is less explicit in learner writing in comparison with native writing, and this could be attributed to language development as well as genre difference.

Table 6-28 Bundles with passive verbs in VP-based bundles in Modular Study 2

CEFR-B2	Freq	**CEFR-C1**	Freq
is based on the	5	*it is believed that*	7
are not allowed to	4	*can be divided into*	4
type	2	**type**	2
token	9	**token**	11

By grouping the word combinations on the basis of structural categorisation, we have seen a few interesting developmental patterns in terms of the use of lexical bundles across learner proficiencies. Despite the different text types or genres investigated in Modular Studies 1 and 2, a rough comparison also suggests that some learner idiosyncrasies are fairly persistent regardless of proficiency levels or genre types. More comparisons across these two studies will be conducted in Chapter 7. In the next section, we turn to functional analysis between CEFR-B2 and CEFR-C1.

6.5 Functional Analysis

In the functional analysis, the distribution of three major functional categories (referential expressions, stance bundles, and discourse organisers, cf. Section 4.3) was investigated. Type and token distribution was distinguished as well. In addition, the chi-square test would likewise be calculated to attest the statistical significance, and standardised residuals would be used to gauge the contribution of each category.

6.5.1 Type Distribution

As can be seen from Table 6-29 and Figure 6-6, discourse organisers take the greatest proportion in both CEFR-B2 writing (46.5%) and CEFR-C1 writing (40.5%), while

referential expressions are the second largest category in both groups of learner writing, 36.6% and 37.8% respectively. With conveying stance as the least frequent bundle function, the ordering of reliance of bundle functions appears to perfectly match that in the two groups of student writing in Modular Study 1 (cf. Section 5.4.1), which will be discussed in detail in Chapter 7.

Table 6-29 Functional distribution in Modular Study 2 (types)

χ^2=0.495, df=2, p=0.781			Function			Total
			Referential expressions	Stance bundles	Discourse organisers	
Corpus	CEFR-B2	Count	26	12	33	71
		% within Corpus	36.6%	16.9%	46.5%	100.0%
	CEFR-C1	Count	14	8	15	37
		% within Corpus	37.8%	21.6%	40.5%	100.0%

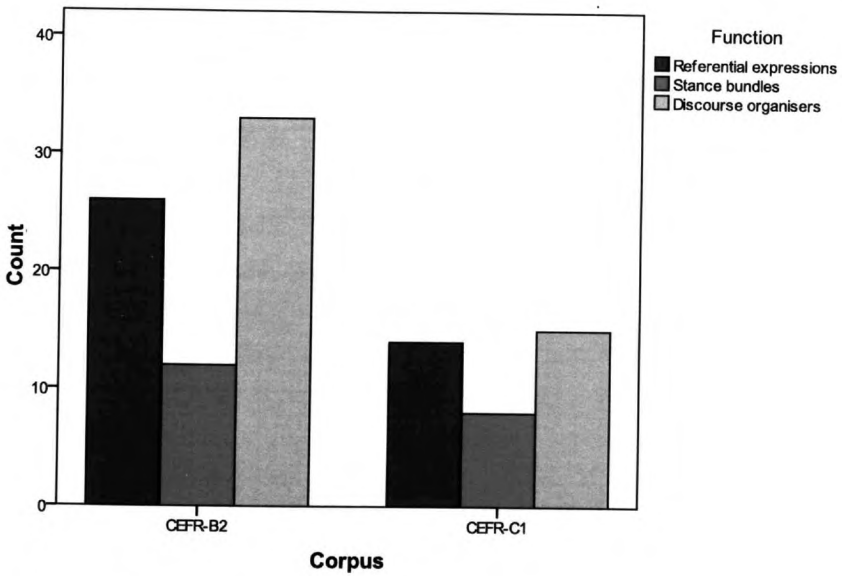


Figure 6-6 Functional distribution in Modular Study 2 (types)

The chi-square test indicates that there is no significant difference at the 0.05 level in terms of functional distribution of bundle types between CEFR-B2 and CEFR-C1 writing with an exceptionally high p value 0.781 (χ^2=0.495, df=2).

6.5.2 Token Distribution

The token distribution of functions in the two subcorpora is similar to the type distribution. As can be seen in Table 6-30 and Figure 6-7, the proportion of discourse organisers remains the highest one in CEFR-B2 (45.7%) and in CEFR-C1 (48.1%), although this proportion increases markedly in CEFR-C1 from 40.5% in the type distribution to 48.1% in the token distribution while it does not fluctuate as much in this regard in CEFR-B2. Referential expressions continue to be the second largest functional category, and stance bundles are still the smallest category in both CEFR-B2 and CEFR-C1 learner writing as in the type distribution.

Table 6-30 Functional distribution in Modular Study 2 (tokens)

χ^2=5.315, df=2, p=0.07			Function			Total
			Referential expressions	Stance bundles	Discourse organisers	
Corpus	CEFR-B2	Count	168	55	188	411
		% within Corpus	40.9%	13.4%	45.7%	100.0%
	CEFR-C1	Count	80	45	116	241
		% within Corpus	33.2%	18.7%	48.1%	100.0%

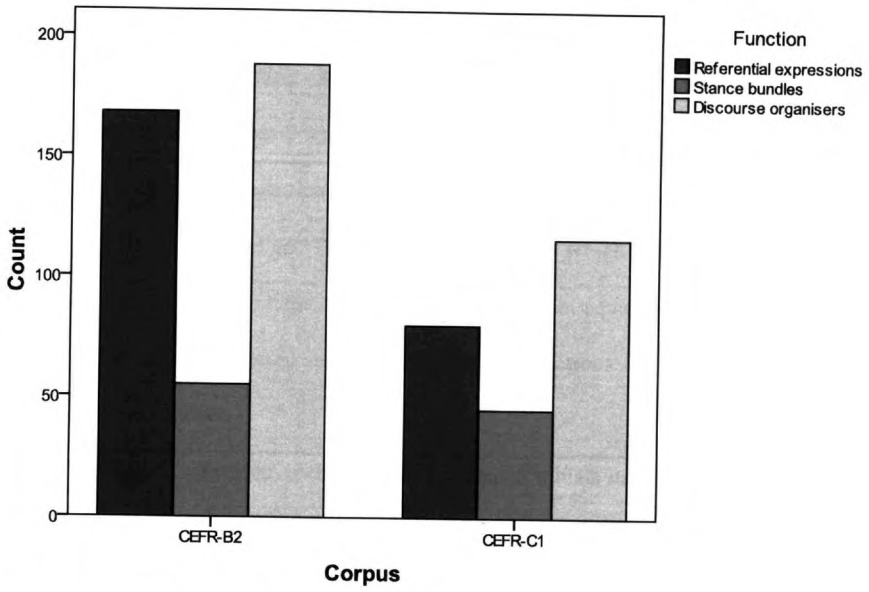


Figure 6-7 Functional distribution in Modular Study 2 (tokens)

The chi-square test indicates that there is no significant difference in terms of functional distribution of bundle tokens between CEFR-B2 and CEFR-C1 writing with the *p* value 0.07 (χ^2=5.315, df=2), a value which nearly reaches the threshold of set significance level 0.05.

Interestingly, token distribution has always shown a higher degree of difference than type distribution in either functional analysis or structural analysis, although there is no significant relationship found between functional distribution between CEFR-B2 and CEFR-C1 writing in both bundle types and tokens. Despite the insignificant relationship, in the next section, we can still make use of the chi-square residuals to examine how these two learner subcorpora distinguish from each other quantitatively in the use of bundle functions.

6.5.3 The Chi-square Test and Standardised Residuals

The rationale of calculating chi-square statistics and standardised residuals has been explained earlier in Chapter 5 and also in this chapter for structural analysis (Section 6.4.3). The identical procedure was executed with the functional categorisation of CEFR-B2 and CEFR-C1 data in this section. The focus would fall on which cells in the two (corpus) by three (functional category) table contribute more to the chi-square statistics so that the extent to which the learners differed in the use of bundle functions between these two proficient groups can be identified.

Table 6-31 shows the results of standardised residuals in the six cells in the matrix of functional distribution in terms of bundle types. As there is no significant difference in functional type distribution, the values in the six cells are all minimal, with the absolute value ranging from 0.1 to 0.4.

Table 6-31 Chi-square standardised residuals for functional distribution (types) in Modular Study 2

χ^2=0.495, df=2, p=0.781		***Referential expressions***	***Stance bundles***	***Discourse organisers***
CEFR-B2	Observed Count (O_i)	26	12	33
	Expected Count (E_i)	26.3	13.1	31.6
	R	-0.1	-0.3	0.3
CEFR-C1	Observed Count (O_i)	14	8	15
	Expected Count (E_i)	13.7	6.9	16.4
	R	0.1	0.4	-0.4

With reference to token distribution, the result shown in Table 6-32 reveals that the two greatest magnitudes, 1.3 and -1.2, were found in the cells representing stance bundles and referential expressions in CEFR-C1 writing, and another two cells representing the same two functional categories in CEFR-B2 also make some contribution to the difference in the statistical test.

Table 6-32 Chi-square standardised residuals for functional distribution (tokens) in Modular Study 2

χ^2=5.315, df=2, p=0.07		Referential expressions	Stance bundles	Discourse organisers
CEFR-B2	Observed Count (O_i)	168	55	188
	Expected Count (E_i)	156.3	63.0	191.6
	R	0.9	-1.0	-0.3
CEFR-C1	Observed Count (O_i)	80	45	116
	Expected Count (E_i)	91.7	37.0	112.4
	R	-1.2	1.3	0.3

The result in the token distribution suggests that the categories of referential expressions and stance bundles may better distinguish groups of learner writing with different proficiency levels in the context of argumentative and expository writing. Recall that the results in Modular Study 1 indicate that discourse organisers and referential expressions are better indicators for writing competency between native professional writing and student writing. Although the association between two studies is not strong, it should be kept in mind that the chi-square test does not indicate significant difference in terms of functional distribution in these two CEFR groups; therefore, the values in the matrix cells in Table 6-31 and Table 6-32 are overall comparably slim.

The original purpose of calculating chi-square standardised residuals is to find out which categories can be used as a better indicator of writing competency. Although the analyses do not appear to pinpoint a good indicator which can serve comprehensively for all sorts of written samples produced by various groups of writers, such a methodology still provides some insight into the quantitative differentiation of the groups compared. This way of carrying out quantitative analysis surely hinges on variables such as quality of data, bundle categorisation, rating procedure, or any other methodological issues that can be thought of, which will all be thoroughly discussed in Chapter 7. Meanwhile, in the next section, the difference of bundle use in each functional category between CEFR-B2 and CEFR-C1 writing will be addressed, which may provide some explanation to the quantitative difference or similarity here.

6.5.4 Use of Lexical Bundles in Functional Categories

Three major functional categories and nine subcategories have been discussed in Section 4.3. In this section, the marked difference and/or similarity will again be discussed in each subcategory. In addition, it has to be noted that only the graphs of type distribution in the breakdown of each functional category are presented as type and token distribution are often quite similar, and the latter usually only shows a greater extent of variation than the former.

6.5.4.1 Referential Expressions

Referential expressions carry three discourse functions: framing, quantifying, and time/place/text deixis. Remember that great reliance on referential expressions, particularly framing bundles, is a characteristic in native academic writing (cf. Section 5.4). As can be seen in Figure 6-8, in EAP-like learner writing represented by CEFR-B2 and CEFR-C1 writing here, such reliance was not detected. The weaker group CEFR-B2, instead, demonstrated a striking dependence on quantifying bundles, nearly one quarter of all its bundle types. Conversely, the more mature writing in CEFR-C1 appear to be more restrained in the use of quantifying bundles as there are only four referential quantifying bundles.

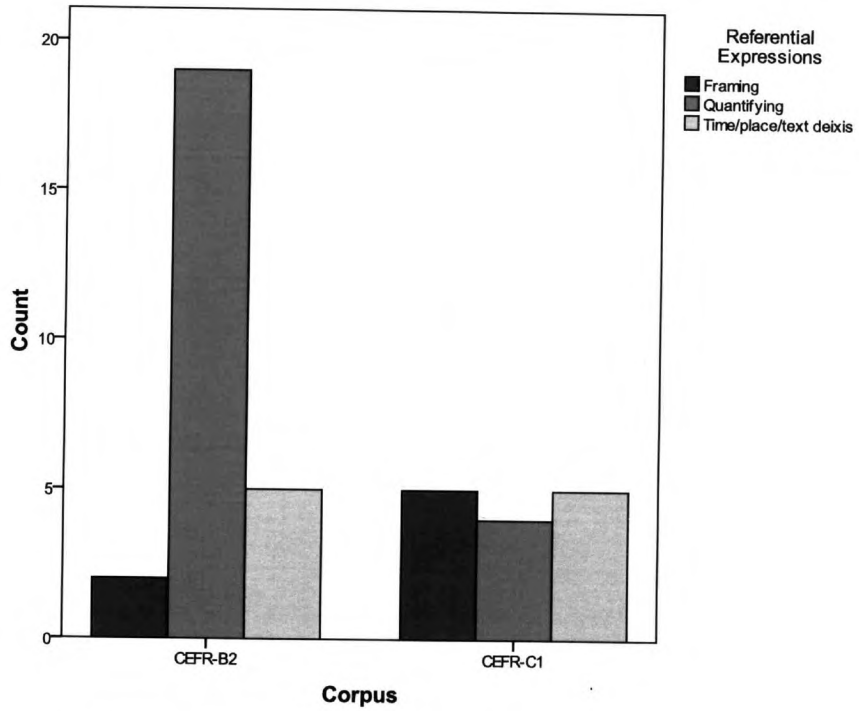


Figure 6-8 Breakdown of referential expressions in Modular Study 2 (types)

From Table 6-33, we can see that nine out of the nineteen quantifying bundles in CEFR-B2 writing subsume the quantifier *a lot of*. Since *a lot of* can be used for both countable and uncountable nouns, it appears to be an omnipotent quantifier for L2 learner writers. However, the quantifying bundles in CEFR-B2 writing also contain other quantifiers such as *a great number of* and *a large amount of*, which indicates that learners at this stage might have already begun to seek alternatives for the superfluous *a lot of*. It is also likely that learners at this level simply quantify excessively with various forms in their writing, and this tendency dies down when they advance to the C1 level.

Table 6-33 Referential quantifying bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
a great number of	5	*a great deal of*	5
a large amount of	4	***some of them are***	4
a lot of people	10	*the rest of the*	6
a lot of problem(s)	16	*there are still some*	4
a lot of time	9		
all of them are	4		
and a lot of	6		
become more and more	6		
bring a lot of	4		
has a lot of	4		
have a lot of	4		
most of the people	5		
most of them are	7		
some of them are	4		
the rest of the world	4		
there are a lot of	11		
there are quite a+(lot of)	5		
there are so many	4		
there are too many	5		
type	19	**type**	4
token	117	**token**	19

However, with a quick scan of the concordance listings, a number of awkward or incorrect collocations were spotted as below:

- ***a large amount of** different culture[39] (CEFR-B2)
- ?**a large amount of** public transport (CEFR-B2)
- ?**a great number of** food and necessity (CEFR-B2)

Errors at the clause level, on the other hand, were found in the bundles *there are so many* and *there are too many*. The existential '*there is/are* + NP' structure appears to be

[39] The asterisk * is used before an example retrieved from the learner writing to indicate the bundle in question is erroneous within the context, and the question mark ? is used when the accuracy or appropriacy of the bundle is questionable within the context. In addition, the problematic part of the sentence is underlined.

commonly misused by L1 Chinese learners of L2 English, probably in the light of the influence of mother tongue. In Mandarin Chinese, there is a similar structure with '有 (yǒu[40], *there is/are*) +NP', which allows a verb phrase to follow and to modify the noun phrase. The learners seemed to be struggling with the distinction of such a similar structure in L1 Chinese and L2 English. Consequently, they simply generated various types of mistakes in the existential structure '*there is/are* + NP' in English, such as placing an infinitive, a finite verb or a relative clause after the NP as demonstrated by the examples below. The point here is that learners repetitively made similar mistakes and increased the occurrences of certain word sequences to the extent that they have been picked up by the computation as one of the defined lexical bundles. Such a pattern of mistakes, however, cannot be disclosed without looking at the concordance lines.

- *The blind have no choice to do other kind of job because ***there are too many*** companies refuse to hire them. (CEFR-B2)
- *From the statistic and information, we can see that ***there are too many*** private cars and which cause traffic congestion. (CEFR-B2)
- ?Everyday ***there are so many*** passengers and goods transport from China mainland to Hong Kong through this way. (CEFR-B2)
- *Why ***there are so many*** prostitutes exists in our society. I think that is because men don't regard women. (CEFR-B2)

The tendency of overstatement in learner writing, as discussed in Section 6.4.4.1, might also have given rise to the supernumerary use of quantifiers in the bundles. Quantifiers found in CEFR-B2 bundles include indefinite pronouns like *all*, *most*, and *some* as well as adjective phrases such as *a lot of*, *a great number of*, *too many* or *so many*. Less proficient

[40] Here 'yǒu' is the Romanised script of the Chinese character '有' with the use of Mandarin phonetic transcription system called *Hanyu Pinyin*.

learners seem to excessively overuse these expressions to reinforce their argument.

In terms of referential framing bundles in CEFR-B2 and CEFR-C1 (see Table 6-34), nothing particular was found, except that they are notably fewer than those in academic writing. As for the third referential subcategory (see Table 6-35), time/place/text deixis, most of the bundles are composed of prepositional phrase fragments, and four of them are shared between these two CEFR-defined groups, which have been discussed in the structural analysis (cf. Section 6.4.4.1).

Table 6-34 Referential framing bundles in Modular Study 2

CEFR-B2	Freq	**CEFR-C1**	Freq
the quality of the	4	*in such a way+(that)*	5
with the development of	6	*in the process of*	6
		on the basis of	4
		the quality of the	6
		the relationship between the	4
type	2	**type**	5
token	10	**token**	15

Table 6-35 Referential deictic bundles in Modular Study 2

CEFR-B2	Freq	**CEFR-C1**	Freq
all over the world	5	***all over the world***	5
at the same time	17	*at the beginning of (the)*	5
for a long time	11	***at the same time***	14
in the following paragraphs	4	***for a long time***	6
the end of the	4	***the end of the***	6
type	5	**type**	5
token	41	**token**	36

6.5.4.2 Stance Bundles

Stance bundles are used to convey the degree of epistemic certainty or probability (epistemic) or the writer's attitude to a proposition (attitudinal/modality). In Figure 6-9, we see very few stance bundles in CEFR-B2 and CEFR-C1, only six and four in each of the subcategories respectively.

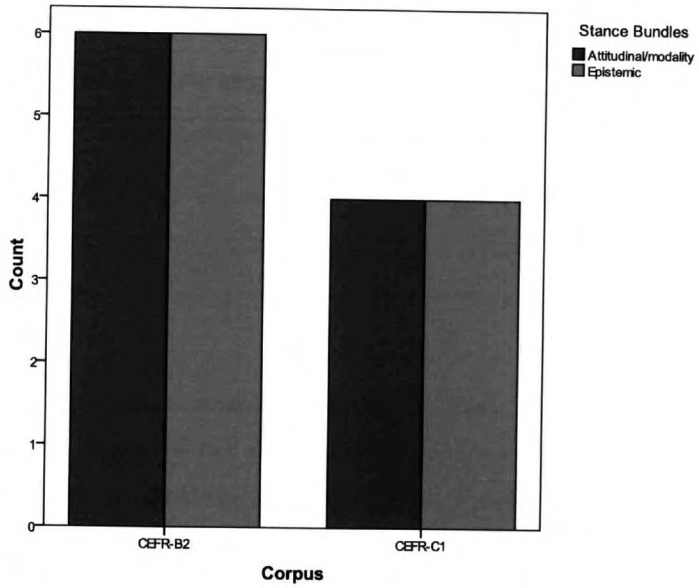


Figure 6-9 Breakdown of stance bundles in Modular Study 2 (types)

The stance epistemic bundles are presented in Table 6-36. As we have seen in Modular Study 1, L2 university students showed a very limited range of expressions used to mitigate the impact of a statement; hence, it is not surprising that no hedging device was found in epistemic bundles in these two groups of learner writing, probably with only one subtle exception, *it is believed that*, in CEFR-C1.

Table 6-36 Stance epistemic bundles in Modular Study 2

CEFR-B2	Freq	**CEFR-C1**	Freq
as a matter of fact	4	***as a matter of+(fact)***	5
as we all know	4	*it is believed that*	7
I think it is	4	*it is obvious that+(the)*	11
I think that this	4	***it is true that***	6
it is true that	4		
some people think that+(the)	4		
type 6		**type** 4	
token 24		**token** 29	

The expression *it is believed that* is considered to be subtle to some degree because its mitigating effect does not appear as strong as other similar expressions found in academic writing such as *it has been suggested*, *is likely to be*, or *it is possible to*. Yet this extraposition structure with the passive voice still has some objective force which suggests hedging.

- To acquire the first language, ***it is believed that*** correction is not necessary. They will naturally correct the errors after communicating more with others, they will know the correct way of using the language. (CEFR-C1)
- Many investigators concluded that the most significant factors responsible for sexual orientation occur prior to birth. ***It is believed that*** an individual is born with a predisposition of being homosexual, heterosexual, or bisexual. (CEFR-C1)

Aside from the possibly only hedging expression *it is believed that*, the better L2 writing represented by CEFR-C1 still tend to be more hedged compared with CEFR-B2 writing. For example, in the concordance lines incorporated with another epistemic bundle *it is true that*, two instances found in CEFR-C1 combined it with *to some degree* or *to some extent* to lessen the impact of the statement, but none of these occurrences in CEFR-B2 did so.

- Yes, ***it is true that*** capital punishment is to some extent effective in some countries but if the question of morality is raised, then this form of punishment is totally immoral. (CEFR-C1)
- They also imitate other people's speech in second language. Yet ***it is true that*** to some degree they may try out constructions. (CEFR-C1)
- ***It is true that*** we will use the language once it is in need. However, these chances never come. (CEFR-B2)

By contrast, one bundle which sounds quite categorical is *as we all know*, only found in CEFR-B2 writing. The weaker learners seemed to simply make a bold assumption that all

the readers know the 'facts' under discussion, which is not necessarily the case. It is speculated that making such brave statements without any qualification probably involves cognitive maturity as well as second language development. As the Longman Learners' Corpus did not record the learners' age or education backgrounds, however, it is impossible to attest this speculation.

- ***As we all know*** now the speed of economy's increasing in China is much faster than the speed in west.
- ***As we all know,*** places in H.K. and Chinese University are inadequate. Therefore, a lot of students study broad after their A. level or High level.

The final noticeable feature discovered in stance bundles is the appearance of the first person pronouns *I* and *we* in CEFR-B2 writing, which has been discussed in the VP-based bundles in the structural analysis (see Section 6.4.4.2). In addition to the bundles with the plural form *as we all know*, there are two other bundles with the first person singular form in CEFR-B2 writing: *I think it is* and *I think that this*. No such bundles were found in CEFR-C1 writing. A concordance search for the word combination *I think* shows 79 occurrences in CEFR-B2 writing and only 29 occurrences in CEFR-C1 writing, which suggests that the weaker group of writing exhibits a higher degree of overt writer visibility than the more proficient group. The issue of author identification will be further discussed in Chapter 7.

Bundles in the other stance subcategory, attitudinal/modality bundles (see Table 6-37), generally express the writer's assessment of a proposition to do something in terms of importance/obligation/difficulty. They are mostly composed of the structures '*it is* + Adj. + *to*' or 'Verb +*to*'. In this regard, general learner writing does not differ much from academic writing as exemplified in the examples given.

Table 6-37 Stance attitudinal/modality bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
are not allowed to	4	*it is hard to*	4
is very important to	5	*it is not easy+(for)*	4
it is difficult to	6	***it is very difficult***	4
it is very difficult+(to)	6	*necessary for us to*	4
should learn how to	4		
will not be able to	6		
type	6	**type**	4
token	31	**token**	16

- I have been reading one of Japanese magazines for more than six years which I am ordering here in Hong Kong because ***it is difficult to*** get Japanese books. (CEFR-B2)
- If your result is just passed, you ***are not allowed to*** go to study Form 6 because there has many students but just a few secondary schools. (CEFR-B2)
- First, young children may easily get confused with two languages and cultures at the same time. ***It is not easy*** for them to accept two different systems, such as spelling, grammar and structure. (CEFR-C1)

6.5.4.3 Discourse Organisers

Discourse organisers are used to introduce a topic, elaborate the topic, identify the focus, or indicate the inferential relationship between the prior and the following text. Figure 6-10 indicates that both CEFR-B2 and CEFR-C1 learners used more discourse organisers for elaboration and identification purposes.

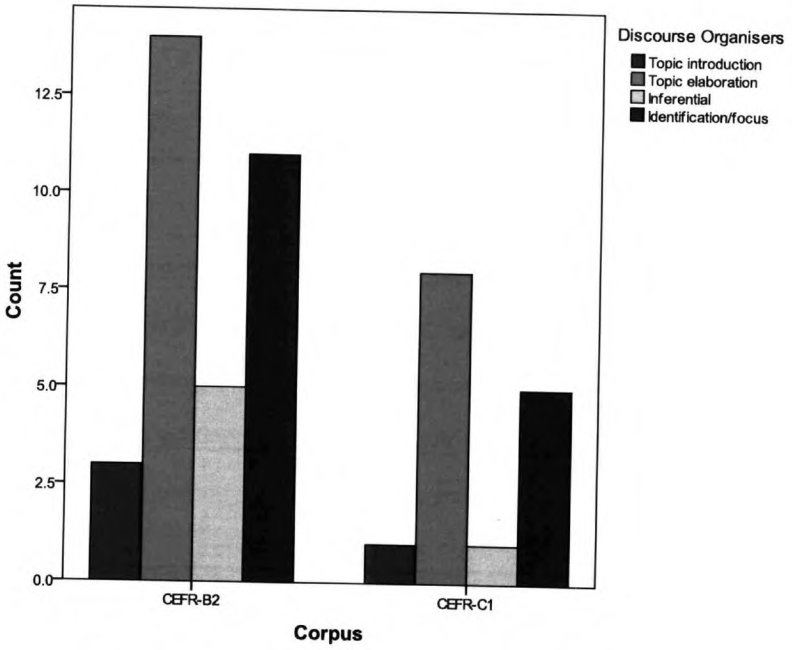


Figure 6-10 Breakdown of discourse organisers in Modular Study 2 (types)

If we start with topic-introduction organisers (see Table 6-38), compare the examples from CEFR-B2 and CEFR-C1 writing below:

Table 6-38 Topic introduction bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
I am going to	4	***I would like to***	11
I would like to	7		
if there is a	4		
type	3	**type**	1
token	15	**token**	11

- In the following paragraph, ***I am going to*** discuss the social problems in Hong Kong associated with urbanisation. (CEFR-B2)
- Lastly, ***I would like to*** mention a couple of points I find unreasonable. (CEFR-B2)

- ***I would like to*** give a brief introduction to my city, Hong Kong about its good and bad features of life. (CEFR-C1)
- In the following paragraphs ***I would like to*** analyze it from both the demand side and supply side and draw a general conclusion in the end. (CEFR-C1)

It is found that the verbs collocated after *I am going to* and *I would like to* in CEFR-B2 writing are largely used to bring up the subject of discussion. They are generally the verb phrases with a semantic meaning that seems more general such as *mention*, *talk about*, or *discuss*. On the contrary, the collocated verb phrases in CEFR-C1 are semantically more specific such as *examine*, *comment*, or *analyze* as shown in Table 6-39.

Table 6-39 Collocation after the topic-introduction bundle in CEFR-B2 and CEFR-C1 writing

Corpus	CEFR-B2	Freq	CEFR-C1	Freq
Topic-introduction	*I am going to*	4	*I would like to*	11
	I would like to	7		
Collocated verb phrase	*write to support*	1	*choose*	1
	point out	2	*examine*	1
	mention	1	*seek*	1
	talk about	3	*pay attention to*	1
	give him some advice	1	*give a brief introduction*	1
	discuss	2	*suggest*	1
	warn him	1	*begin*	1
			use	1
			restrict myself to discussing	1
			comment	1
			analyze	1

Verbosity within discourse organisers was also observed throughout learner writing, particularly in CEFR-B2 writing (see the other three subcategories of discourse organisers in Table 6-40 to Table 6-42). For the purpose of being more concise, examples below can be easily paraphrased with the use of fewer words.

- ***It is a very*** interesting question that why China has the high share if investment in relation to GDP compared with other developing countries. (CEFR-B2)

- In a word, advertising plays ***a very important role*** in our society and helps the economic growth of the country. (CEFR-B2)

- Learning foreign languages, one must concentrate on practising speaking, listening and memorizing. Especially, ***the most important thing*** is practising speaking and listening a lot. (CEFR-B2)

Compared with the learner writers in CEFR-C1, learners in CEFR-B2 tended to use more seemingly redundant discourse organisers in every respect to structure their ideas. One noticeable prevalent use is the structures '*there is/are* + NP' and '*it is* + NP' in topic elaboration/clarification bundles (see Table 6-40, e.g. *but there are still*, *there will be a*, *it is also a*), which leads to a style suggestive of simplicity and verbosity. All the above features indicate that the less proficient learners relied heavily on the explicit lexico-grammatical expressions to organise the text. These explicit discourse organisers more or less result in a simplistic and verbose style which would not be desired in academic writing. On the other hand, the more competent writers in CEFR-C1 coped with the discourse organisers with more sophistication and relied on the discourse organisers to a lesser extent, although the endeavour of organising the text is still visible.

Table 6-40 Topic elaboration/clarification bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
(can)+have the right to	14	*as well as the*	7
and to be a	4	*can be divided into*	4
but there are still	4	***has the right to***	4
has the right to	5	*how to deal with*	4
is based on the	5	*in order to make*	5
is more important than	4	*is a kind of*	5
is totally different from	4	***on the other hand***	28
it is a good	4	*to cope with the*	4
it is a very	4		
it is also a	7		
it is not a	4		
on the other hand	18		
there will be a	4		
want to be a	5		
type	14	**type**	8
token	86	**token**	61

In terms of inferential bundles, there are only a few such instances in both CEFR-B2 and C1 writing (Table 6-41). It is speculated that in order to indicate the relationship between cause and result, learners at these levels might more often resort to common causal words such as *because*, *so*, or *therefore* as opposed to the more sophisticated longer inferential devices as found in Modular Study 1 (e.g. *in the light of*, *in view of the*).

Table 6-41 Inferential bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
as the result of	4	*as a result of*	4
because they are not	4		
the main reason is	5		
the result of the	4		
the result of this	5		
type	5	**type**	1
token	22	**token**	4

In the final subcategory of discourse organisers (see Table 6-42), some of the identification/focus bundles in CEFR-B2 writing are indicative of a hyperbolic tone, e.g. *is the most important*, *the best way to*. Combined with the observations from other

subcategories of discourse organisers (e.g. *is totally different from*, *it is a very*), CEFR-B2 writing can be characterised as having this hyperbolic style whereas this tendency subsides to a very large extent in CEFR-C1 writing.

Table 6-42 Identification/focus bundles in Modular Study 2

CEFR-B2	Freq	CEFR-C1	Freq
(from)+my point of view	5	*as far as the*	4
(is)+the best way to	5	***is one of the***	14
a very important role	4	***one of the most***	11
as I have mentioned	4	*we can say that*	5
him or her to	4	*we can see that*	6
is one of the	18		
is the most important	4		
one of the most	5		
people who live in	5		
the most important thing+(is)	7		
we can see the	4		
type	11	**type**	5
token	65	**token**	40

6.6 Relationship between Structural and Functional Categorisation

The strong association between form and function of lexical bundles has been illustrated in Modular Study 1 in Section 5.5. This interaction is still prominent with the lexical bundles retrieved from rated CEFR-B2 and CEFR-C1 learner writing (see Figure 6-11). As in Modular Study 1, referential expressions remain to be dominated by NP-based and PP-based bundles although the proportion of VP-based bundles slightly increases. The most extreme distribution is found in stance bundles, in which VP-based bundles still take the majority and there is no presence of NP-based bundles. As to discourse organisers, VP-based bundles make up over half of this category and NP-based and PP-based bundles take the rest of this category.

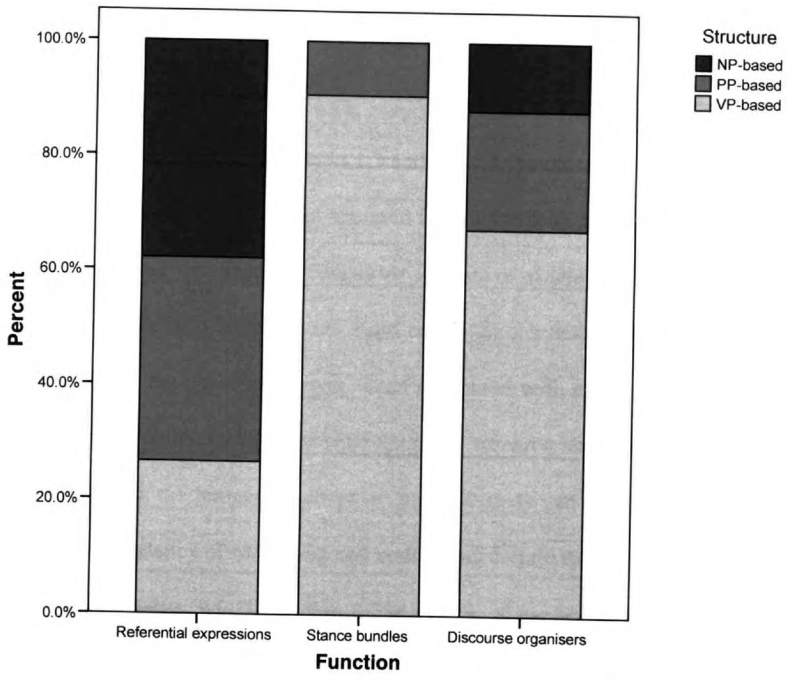


Figure 6-11 Interaction between structural and functional categorisation in Modular Study 2

At the outset of Modular Study 2, it was expected that the more capable writers would use more NP-based bundles and more referential expressions than the weaker writers as revealed in Modular Study 1. However, the statistical tests in this chapter indicate that there is generally no significant difference between CEFR-B2 and CEFR-C1 writing in terms of structural or functional distribution irrespective of bundle types or tokens except for the structural token distribution, in which PP-based bundles made major contributions to the significant difference. Consequently, the strong interaction between forms and functions of lexical bundles does not manifest itself in the use of bundles which allow us to easily distinguish these two groups of writing of different proficient levels.

Although the quantitative analysis in Modular Study 2 turns out to be rather different from expectation, the intense interaction between bundle forms and bundle function has been

consolidated across various designs of studies and different groups of writing.

6.7 Keyness Analysis

As has been discussed in Sections 4.1.3 and 5.6, a keyness analysis applies the same principle as a keyword analysis, but it was renamed in this thesis so as to avoid confusion because this approach is used for word combinations instead of single words. The primary purpose of keyness analysis is to highlight the word combinations that occur significantly frequently or infrequently in the specified corpus when compared with a reference corpus. This approach could be particularly useful in second language research as it can reveal significant overuse and underuse in the learner language so that a remedy can be offered for the L2 learners to rectify their tendency of overusing and underusing certain expressions.

The *p* value, which represents the degree of danger of error, was set at 0.001 as in Modular Study 1. For the first step, CEFR-C1 was set as the reference corpus because it represents a higher proficiency level that the weaker learners in CEFR-B2 would like to achieve. After the key bundles have been generated by *WordSmith 4.0*, the next step was to crosscheck the key clusters retrieved to ensure that the target key bundles met the cut-off frequency and dispersion requirements set in this study, i.e. occurring at least four times in three texts or more. After removing clusters that were not within the defined lexical bundles, the filtered lexical bundles in CEFR-B2 were finalised as shown in Table 6-43 with the key value in the brackets. The higher the key value, the more statistically frequent the bundle in question when compared with CEFR-C1.

Table 6-43 Key lexical bundles in CEFR-B2 with CEFR-C1 as the reference corpus

Corpus	Overuse/underuse	Key lexical bundles
CEFR-B2	Overuse	*have the right to* (19), *there are a lot of* (13), *a lot of people* (10), *a lot of time* (9)

Four key bundles were found in CEFR-B2 writing, which were all overused. The bundle with the strongest keyness is *have the right to*. It also has two other variations found in the CEFR-B2 writing: *has the right to* and *can have the right*, but the latter has been incorporated with *have the right to* into one longer unit *(can)+have the right to* in the final repertoire because of the shared occurrences at the stage of 'cleaning up' overlapping (cf. Section 3.6.2). Most of the instances with the use of *(can)+have the right to/has the right to* occur in the light of the topics in diverse writing tasks including *Feminism*, *Race and Immigration in the UK*, *Homosexuality*, *Jury*, *Learning is Accumulative*, and *Abortion*. Some of the occurrences are as follows:

- Besides the natural characters, the women think they're the same as man. They ought to ***have the right to*** do what they want to do and be able to chase after reputation. (CEFR-B2)
- It is because either the defendant or prosecutor [sic] ***have the right to*** challenge any juror in the trial and demand for the replacement. (CEFR-B2)
- No one ***has the right to*** laugh at me. I know I should not be indifferent or I will lose much from learning. (CEFR-B2)

Although great efforts have been made to exclude context-dependent bundles, the examples above might seem ambiguous. The fact that they do come from various writing tasks is the reason why *(can) have the right to* is considered to be context-independent and kept in the finalised bundles.

In addition to the most overused *have the right to*, the other key bundles all contain *a lot of*, a quantifier with a colloquial and overstating tone. This keyness analysis confirms the overuse of this quantifier in CEFR-B2 writing found earlier in Sections 6.4.4.1 and 6.5.4.1.

6.8 Summary of Findings

Various approaches and perspectives have been adopted with an attempt to present the whole picture of similarity and difference between CEFR-B2 and CEFR-C1 writing. The findings are now grouped and summarised as below:

- as shown in the total columns of Table 6-21, Table 6-22, Table 6-29, and Table 6-30, the number of lexical bundles decreases with the advance of writing competency for both overall frequency (tokens) and the range of lexical bundles (types), which is opposite to the pattern revealed in Modular Study 1 and will be further discussed in Chapter 7;
- In terms of structural distribution, VP-based bundles take nearly half of the bundle population in both CEFR-B2 and CEFR-C1 writing regardless of bundle type or token distribution;
- chi-square tests show that there is a significant difference of distribution in terms of bundle structures between CEFR-B2 and CEFR-C1 writing in token distribution ($p<0.0005$) but not in type distribution ($p=0.12$);
- a further examination of chi-square standardised residuals suggest that PP-based bundles are the major structural category that contribute to the difference of token distribution between CEFR-B2 and CEFR-C1 writing, with CEFR-B2 being the group significantly underusing PP-based bundles ;
- the weaker group, CEFR-B2, used the quantifier *a lot of* in various lexical bundles (around 10% of the total bundle types), which could be due to the learners' tendency towards overstatement, and this tendency can be evidenced by other bundles shared by the two learner writing groups, such as *for a long time*, *all over the world*;

- both learner writing groups show a very limited range of interchangeable nouns which can fit in the noun slot in the two frames, '*the* + Noun + *of/the*' and '*in the* + Noun + *of*';
- there appears to be a shift of authorial tone in terms of the use of personal pronouns in the bundles as the writing competency develops, from the first person singular pronoun *I* to the plural pronoun *we* and then to the impersonal pronoun *it*, and bundles with passive verbs are rare in both CEFR-B2 and CEFR-C1 writing;
- in terms of functional analysis, discourse organisers are the most frequently used bundles, with referential expressions as the second frequently used ones and stance bundles the last, the ordering of which holds true for both CEFR-B2 and CEFR-C1 writing regardless of bundle type or token distribution;
- chi-square tests show that there is no significant difference of distribution in terms of bundle functions between CEFR-B2 and CEFR-C1 writing in either bundle type distribution (p=0.07) or bundle token distribution (p=0.781);
- the weaker group, CEFR-B2, shows a strikingly high reliance on quantifying bundles such as *a large amount of*, *there are too many* (nearly 26% of the total bundle types), and some of the collocations or structures with the quantifying bundles are misused or inappropriate;
- from the investigation of stance bundles, CEFR-C1 writing is found to be more hedged than CEFR-B2 writing, although few explicit hedging devices have been seen in the CEFR-C1 bundle repertoire;
- the verb or verb phrases collocated with the topic-introducing bundles such as *I would like to* in CEFR-B2 carry a semantic meaning which are more implicit (e.g. *point out*, *discuss*, *talk about*) while the collocated verb or verb phrases found in CEFR-C1 have more specific semantic denotation (e.g. *examine*, *suggest*, *comment*);

- CEFR-B2 writing is found to be indicative of a simplistic and verbose style in the use of discourse organisers (e.g. *the most important thing*, *it is a very*, *but there are still*);
- there is a strong relationship between structural and functional categorisation in Modular Study 2: referential expressions are mostly made up of NP-based and PP-based bundles while stance bundles are dominated by VP-based bundles, and discourse organisers show a more balanced distribution between VP-based and NP/PP-based bundles; and
- the keyness analysis confirms the overuse of the quantifier *a lot of* in CEFR-B2 writing.

It has to be emphasised that the findings addressed above were all established upon the learner data which have undergone a complex and robust selection procedure as reported in the first part of this chapter. Without a careful plan and execution of the rating process, it is impossible to conduct bundle analysis and make comparison of these lexical bundles between two groups of learner writing with defined proficiency levels.

Meanwhile, through an investigation of these two CEFR subcorpora, a few developmental patterns have been observed, and some of the learner idiosyncrasies have also been spotted in Modular Study 1, in which L2 learner writing was compared with two groups of native writing in the academic context. It appears that some learner idiosyncrasies persist to some extent despite the genre and proficiency difference while others gradually diminish with proficiency progress. In the next chapter, therefore, an overall comparison will be made between Modular Studies 1 and 2, and we shall see whether these rough observations are empirically valid under a much closer examination.

Chapter 7 Overview & Discussion

This chapter mainly aims to address the three explanatory research questions (see Section 3.1). The results of quantitative and qualitative analyses from Modular Studies 1 and 2 will be first compared to investigate to what extent the five groups of varying writing proficiencies differ from or resemble each other and whether a development pattern can be identified. The structural and functional distributions of lexical bundles will be explored among the five groups of writing. Then the results of keyness analyses as well as the chi-square residuals in Modular Studies 1 and 2 will be compared and discussed. Following the overview of the two Modular Studies, distinctive features disclosed across writing development will be examined from a prescriptive perspective rather than a descriptive one, which illustrates some possible impacts the results have with respect to improving pedagogy for second language writing. This chapter will be rounded off with a discussion which aims to answer the explanatory research questions. First, it proposes some explanations as to what may have resulted in the learner idiosyncrasies found in the use of lexical bundles. Next, it discusses the possible impacts that the analysis results have on language testing, particularly upon the empirical underpinnings of a rating scale.

7.1 Comparison of Analyses

Although the written texts researched in Modular Study 1 focus on academic writing while Modular Study 2 is concerned with general learner writing with defined proficiency levels, it is still considered feasible to compare the use of lexical bundles in the five groups of writing altogether. On the one hand, some linear relationship does seem to emerge on the basis of the use of lexical bundles across writing proficiencies despite genre and task variation. On the other hand, we will see that certain expressions tend to occur exclusively in learner writing (learner bundles) or in native writing (native bundles) while a number of word combinations

appear to be indicative of academic writing (academic bundles).

Additionally, in the previous two chapters we have seen that bundle structures and bundle functions are closely associated. Due to the constraint of space, I decided to focus on the qualitative examination of discourse functions rather than discussing both structures and functions in detail after the quantitative analyses. Yet bundle structures are still considered in the discussion of distinctive features addressed in Section 7.2.

7.1.1 Quantitative Analyses

For structural distribution of lexical bundles, the native expert writing in FLOB-J shows the highest percentage of use of NP-based bundles and the lowest percentage of use of VP-based bundles while the weakest writing represented in CEFR-B2 exhibits the highest percentage of VP-based bundles and the lowest percentage of PP-based bundles (see Figure 7-1 and Figure 7-2). If we focus on the three groups of learner writing only, CEFR-B2, CEFR-C1, and BAWE-CH, then the pattern is rather consistent. The least proficient CEFR-B2 writing shows the highest proportions of VP-based and NP-based bundles and the lowest proportional use of PP-based bundles. As writing proficiency progresses, the use of VP-based and NP-based bundles decreases whereas the use of PP-based bundles increases. This pattern holds true for both the type and token distributions. The possibly strongest learner writing in BAWE-CH, hence, demonstrates the highest percentage of use of PP-based bundles and the lowest percentage of use of NP-based and VP-based bundles among learner groups. There appears to be a salient linear relationship between learner proficiency and proportion of use of bundle structures.

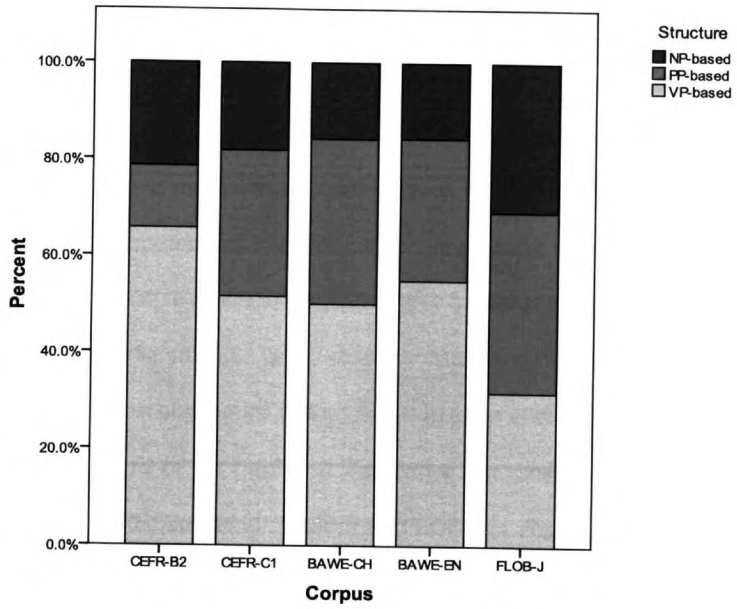


Figure 7-1 Overall structural distribution (types-percentage)

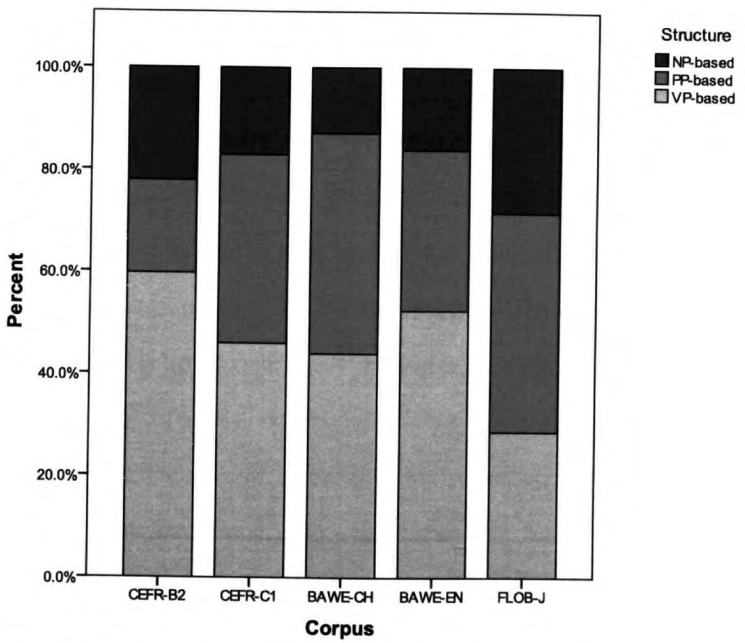


Figure 7-2 Overall structural distribution (tokens-percentage)

It has to be noted that the British student writing represented in BAWE-EN, however, does not look to be positioned in the same ability continuum between the three groups of L2 English and native expert writing. In addition, the expectation that the native British student writing would be most similar to native expert writing (due to both groups having English as a shared mother tongue) is not borne out. Instead, the Chinese peer student writing seems closer to the native expert writing in terms of bundle structure proportions. More quantitative evidence will be provided in this chapter to support this initial observation, and we will also see whether this observation still holds true in terms of the actual use of individual bundles.

In terms of the functional distribution, the patterns do not come into existence as explicitly as the structural distribution. FLOB-J is distinct from the other four groups of comparably immature writing in the sense that it shows the highest percentage of use of referential expressions and the lowest percentage of use of discourse organisers. This pattern in the native expert writing holds true in both the type and token distributions (see Figure 7-3 and Figure 7-4). Aside from FLOB-J, there does not appear to be a clear relationship between the proportions of discourse functions and writing proficiency. In the type distribution (Figure 7-3), the proportion of referential expressions slightly increases with learner proficiency from CEFR-B2 to BAWE-CH while the proportional use of the other two discourse functions fluctuates. Then as can be seen in the token distribution (Figure 7-4), the percentage of referential expressions in CEFR-B2 writing increases to a certain extent and thus violates the only potential pattern found. In the qualitative analysis in the next section, we will see that the increase of referential expressions in this weakest writing group is actually the result of learners' tendency towards overzealous overstatement, embodied in the repetitive use of certain speech-like quantifying expressions. For future research, it might be worth distinguishing between colloquial bundles and academic bundles before making comparisons so as to seek a better defined linear relationship between bundle use and proficiency.

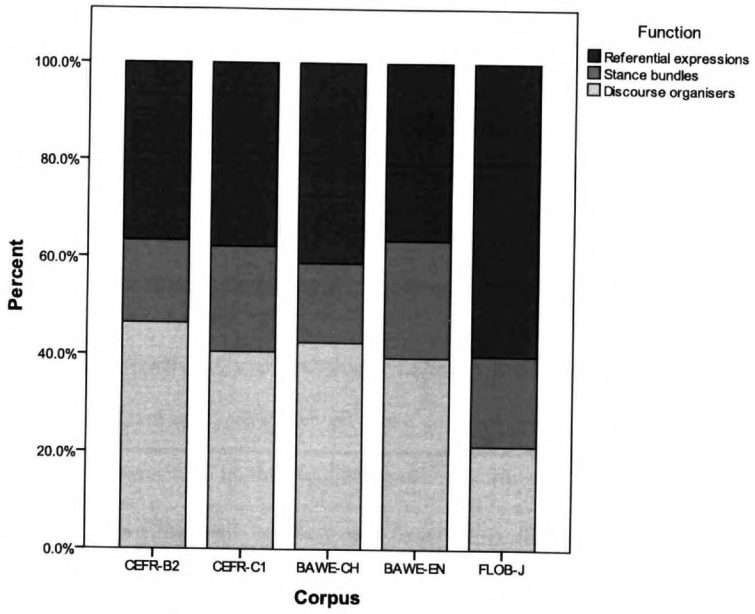


Figure 7-3 Overall functional distribution (types-percentage)

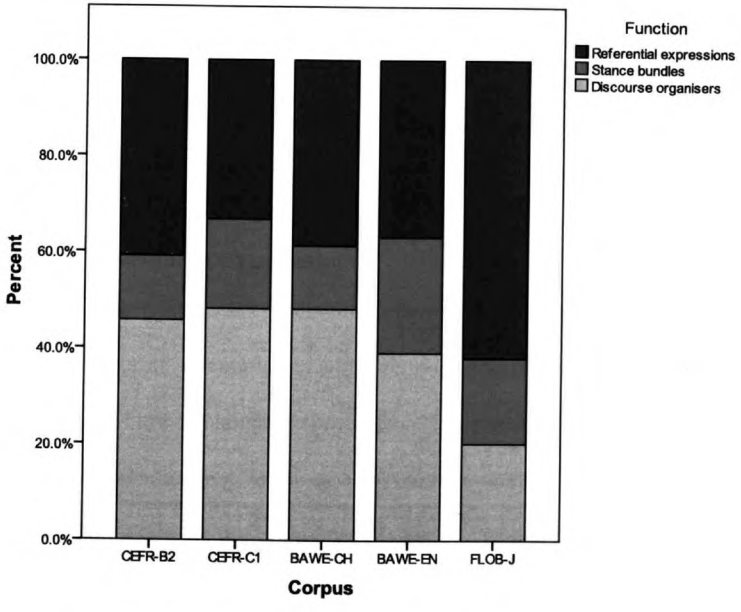


Figure 7-4 Overall functional distribution (tokens-percentage)

As to the British student writing represented in BAWE-EN, again, it does not resemble native expert writing any more than their Chinese peer student writing, except for showing a slightly lower percentage of discourse organisers. Yet bear in mind that in the qualitative analysis of Modular Study 1, the native student writing (BAWE-EN) was found to be more qualitatively similar to the native expert writing (FLOB-J) than quantitatively in some aspects, which will be further illustrated in the following sections.

7.1.2 Qualitative Examination of Discourse Functions

The quantitative analyses above give us a quick overview of the use of lexical bundles across the five subcorpora. In this section, more specific pragmatic/discourse features in distinct groups of writing will be disclosed by scrutiny of functional subcategories and lexical bundles shared in certain subcorpora but not others.

It also has to be stressed that the discussion drawn from quantitative comparisons among subcorpora in this section are primarily established upon proportional data as shown in the graphs as opposed to raw frequency for the sake of comparability. Yet we have to bear in mind that sometimes the comparison is made between 15 discourse organisers in CEFR-C1 and 34 discourse organisers in BAWE-CH, and hence the results should be interpreted with great caution.

7.1.2.1 Referential Expressions

Starting with referential expressions, Figure 7-5 and Figure 7-6 illustrate the proportions of sub-functions of referential bundles in each corpus, which can be utilised to reflect the relationship between language proficiency and referential functions. We can see that the use of framing bundles (e.g. *in the case of*) tends to increase towards the most competent writers on the right while the use of quantifying bundles (e.g. *a large amount of*) appears to decrease. The highest proportion and thus most noticeable point is the use of quantifying expressions in CEFR-B2 writing.

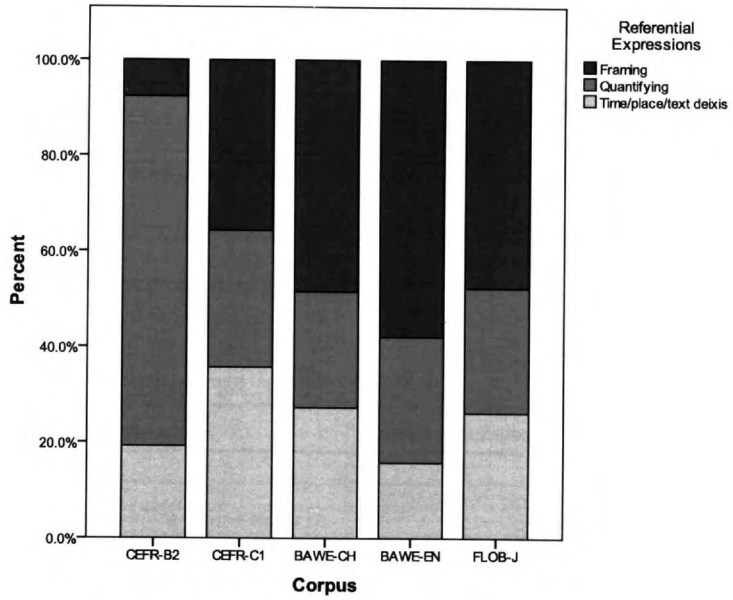


Figure 7-5 Distribution of referential subcategories (types-percentage)

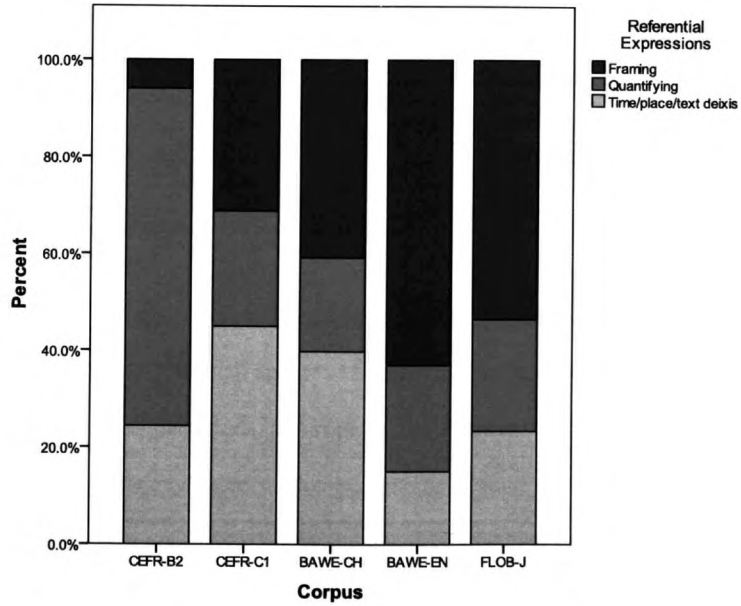


Figure 7-6 Distribution of referential subcategories (tokens-percentage)

Referential expressions in this lower proficiency group can be characterised by the drastic reliance on quantifiers (Table 7-1), particularly the colloquial and overstating cluster *a lot of*, which did not occur in the bundle repertoire in any other subcorpora compared. The overzealous intention to impress the reader(s) with an overstated tone appears to be one of the distinct features in the less proficient learner writing, and this hyperbolic tone attenuates when learners progress to the CEFR-C1 level.

Table 7-1 Referential quantifying bundles in CEFR-B2

Quantifying structure		**Lexical bundles***
Quantifier *a lot of*	*a lot of* + Noun	*a lot of people*$_{(10)}$, *+a lot of problem(s)*$_{(16)}$, *a lot of time*$_{(9)}$, *and a lot of*$_{(6)}$
	Verb + *a lot of*	*bring a lot of*$_{(4)}$, *has a lot of+*$_{(4)}$, *have a lot of+*$_{(4)}$
	there are + a lot of	*there are a lot of+*$_{(11)}$, *there are quite a+(lot of)+*$_{(5)}$
Other quantifiers	*a* + quantifier *+of*	*a great number of*$_{(5)}$, *a large amount of*$_{(4)}$
	there are	*there are so many*$_{(4)}$, *there are too many*$_{(5)}$
Determiner pronouns		*all of them are*$_{(4)}$, *most of the people*$_{(5)}$, *most of them are*$_{(7)}$, *some of them are*$_{(4)}$
Others		*become more and more*$_{(6)}$, *the rest of the world*$_{(4)}$

*The frequency is indicated in the brackets.

If the above overall amplifying quantifiers are compared with those discovered in Modular Study 1, they all appear to be quite general and vague. On the other hand, the degree of quantity or magnitude involved in the recurrent quantifying expressions in academic writing are largely more moderate, and the notions involved are more specific, too. For example, there are several referential quantifying bundles containing *a number of* in Modular Study 1 (e.g. *has a number of*, *in a number of*). Specific nouns were also used in expressions such as *a wide range of* and *a high level of* in Modular Study 1. This distinction might result

from the fundamental difference between the genres of texts, academic writing and general argumentative/expository writing, as the former has to be more precise. It might also be concerned with the competency of writers. As the writers become more mature and proficient, they tend to be more cautious in the use of quantifying expressions.

Remember that the BAWE-EN writing has often displayed an idiosyncratic pattern when quantitatively aligned with learner writing and native expert writing. In Figure 7-5 and Figure 7-6, it has also been noted that the proportion of referential framing bundles in BAWE-EN is slightly higher than that in native expert writing represented in FLOB-J. After a careful examination of referential expressions in BAWE-EN, the preference of *use* in framing bundles (*and the use of*, *for the use of*, *the use of the*, *through the use of*) was found to attribute to this unusually high proportion of framing bundles, which has been discussed in Chapter 5.

7.1.2.2 Stance Bundles

Turning to stance bundles, there appears to be a boundary that distinguishes academic writing and EAP-like learner writing (see Figure 7-7 and Figure 7-8). Overall speaking, the use of epistemic bundles (e.g. *it is believed to*) increases with the progress of writing proficiency whereas the use of attitudinal/modality bundles (e.g. *it is difficult to*) decreases. This linear relationship is particularly pronounced in the token distribution (see Figure 7-8). Yet the generalisations have to be tackled with caution as there are 55 stance bundle tokens in CEFR-C1 and 45 in CEFR-B2 writing while FLOB-J has as many as 125 stance bundle tokens.

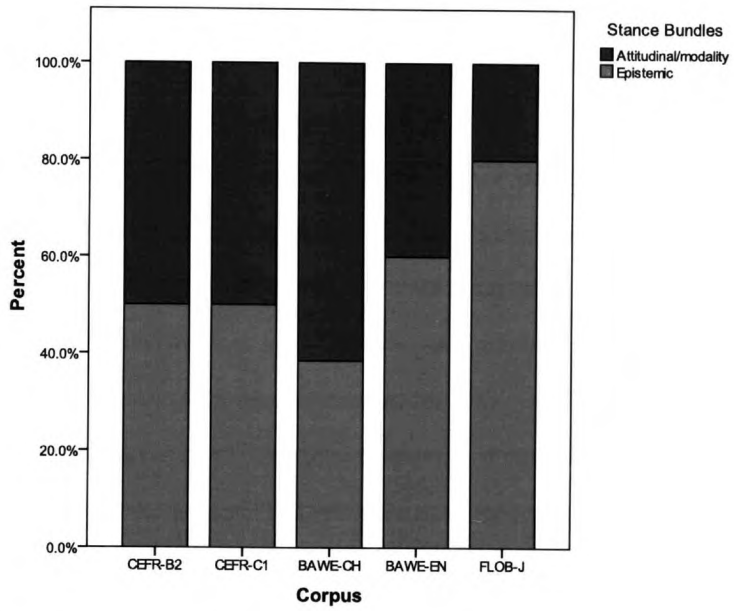


Figure 7-7 Distribution of stance subcategories (types-percentage)

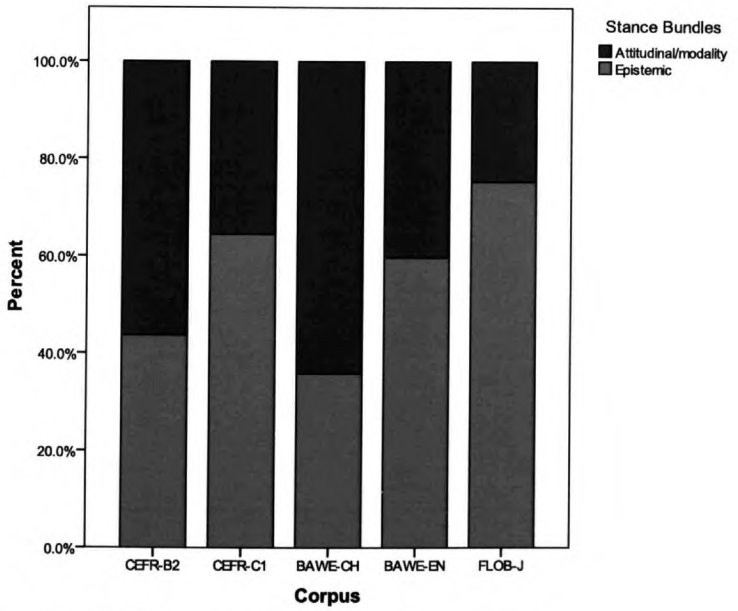


Figure 7-8 Distribution of stance subcategories (tokens-percentage)

As discussed earlier, hedging is generally embodied in the form of epistemic bundles. Looking more carefully into the epistemic bundles used in the five groups of writing, a scale of certainty illustrated with epistemic bundles was constructed, in which the personally involved epistemic bundles and bundles with a strong, medium, and weak degree of commitment are roughly grouped (Table 7-2). As can be seen, as we move towards the least proficient writing of CEFR-B2, the stance tends to be more categorical, arbitrary, and personally involved (e.g. *as we all know*, *as a matter of fact*, *it is true that*). On the other hand, claims tend to be more cautious and carefully worded (e.g. *it has been suggested*, *it is possible to*, *are more likely to*) towards the most proficient writing of FLOB-J.

It should be noted that this scale was established by my own judgment; therefore, some of the determinations might be subjective and disputable. However, at least it appears that a rough continuum of epistemic expression with varying degree of certainty can be charted along writing development.

Table 7-2 Epistemic bundles with the gradient of certainty

Gradient of Certainty	Corpus (word count) / Epistemic bundles Freq	Modular Study 2		Modular Study 1		
		CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
personalised	as we all know	4				
	I think it is	4				
	I think that this	4				
	some people think that (the)	4				
Strong	as a matter of fact	4	5			
	to the fact that			4		5
	is the fact that				6	
	the fact that the				8	
	the fact that they				4	
	the fact that this					4
	by the fact the					9
	it is true that	4	6			
	is by no means				4	
Medium	it is obvious that (the)		11			
	it is believed that		7	5		
	it is clear that				8	11
	it is not clear					4
	there is no evidence				4	
	there is evidence that					5
	to a large extent					4
	it can be argued+(that)				5	
	it could be argued that+(the)				14	
Weak	would have to be				6	8
	would need to be				4	
	would be difficult to					5
	it would have been				5	
	seems to have been					6
	to a certain extent				4	
	was not so much					4
	is considered to be			4		
	it has been suggested (that)			6		4
	it is estimated that				4	
	it is possible to				13	6
	are likely to be				6	4
	are more likely to			5	6	7
	is likely to be					7
	whether or not to					5

It has also been observed from Table 7-2 that active voice constructions are mostly used by CEFR-B2 writers (e.g. *I think it is*) while at the higher proficiency levels, the passive and the impersonal constructions are preferred (e.g. *is considered to be*, *it is clear that*, *there is evidence that*). A retrieval of the use of first person pronouns in the lexical bundles (see Table 7-3) consolidates this preliminary observation. The first person singular is found to be favoured by the less proficient writers. In terms of the function of the first person plural, the learner writers tend to use the bundles with *we* as discourse organisers to identify the proposition to come (e.g. *we can see that*) while the native expert writers tend to refer to *we* to guide the readers for referential deictic function (e.g. *as we have seen*, *as we shall see*).

Table 7-3 Use of first person pronouns in lexical bundles

Epistemic bundles	Freq \ Corpus (Word count)	Modular Study 2		Modular Study 1		
		CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
I/my	(from)+my point of view	5				
	as I have mentioned	4				
	I am going to	4				
	I think it is	4				
	I think that this	4				
	I would like to	7	11			
	in this essay I				4	
we	as we all know	4				
	we can see the	4				
	we can see that		6	7		
	we can say that		5			
as we	as we have seen					8
	as we shall see					7

In addition to the word combinations incorporated with hedges and first person pronouns, another two kinds of lexical bundles from the two modular studies have been identified which might also relate to voice construction: extent/mode modifiers and

overstating bundles. As has been discussed in Chapter 5, the former are word combinations used to modify the extent or manner of a proposition (e.g. *the degree to which*, *in so far as*), and they only occur in native writing FLOB-J and BAWE-EN. The latter are those word combinations with an over-generalising tone (e.g. *all over the world, for a long time*) and, interestingly enough, they are only found in learner writing BAWE-CH, CEFR-C1 and CEFR-B2. According to the *Voice Intensity Rating Scale* devised by Helms-Park and Stapleton (2003), four rhetorical categories conveying authorial voice can be distinguished: *assertiveness*, *self-identification*, *reiteration of central point*, and *authorial presence and autonomy of thought*. We have seen that these frequency-driven phraseological units also contribute to the construction of individualised voice. The lexical bundles discussed above expressing hedging or extent/mode modifying hence belong to the category of assertiveness while those incorporated with first person pronouns, passive or impersonal pronouns fall in the category of self-identification. The ways in which the authorial voice is constructed appears to differ across writing proficiency, which has great implications for L2 writing instruction. The results of comparisons above suggest that weaker writers tend to demonstrate stronger commitment to their assertions and identify themselves more frequently with the use of first person singular pronoun. On the other hand, more competent writes are more cautious in making a strong commitment to their assertions, and they prefer to use the first person plural pronoun for text deictic purposes. They also tend to use the passive or impersonal constructions more frequently than learners. Again, the genre difference also plays an important role here.

7.1.2.3 Discourse Organisers

In this section, we will concentrate on discourse organisers. In Figure 7-9 and Figure 7-10, there does not immediately seem to be as clear a pattern in the distribution of sub-functions in discourse organisers as in referential expressions and stance bundles. If we separate the native

groups from the learner groups, however, it appears that native writers relied on a wider range of inferential expressions and used them more frequently than learner writers. As to the identification/focus expressions, there exhibits a reverse pattern in the sense that native writing shows less reliance on identification/focus expressions than learner writing in terms of both bundle types and tokens.

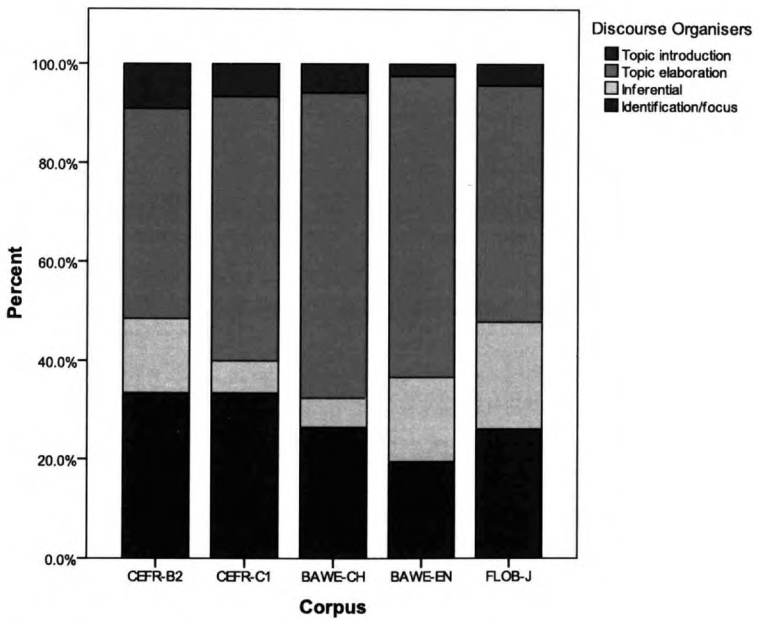


Figure 7-9 Distribution of discourse organising subcategories (types-percentage)

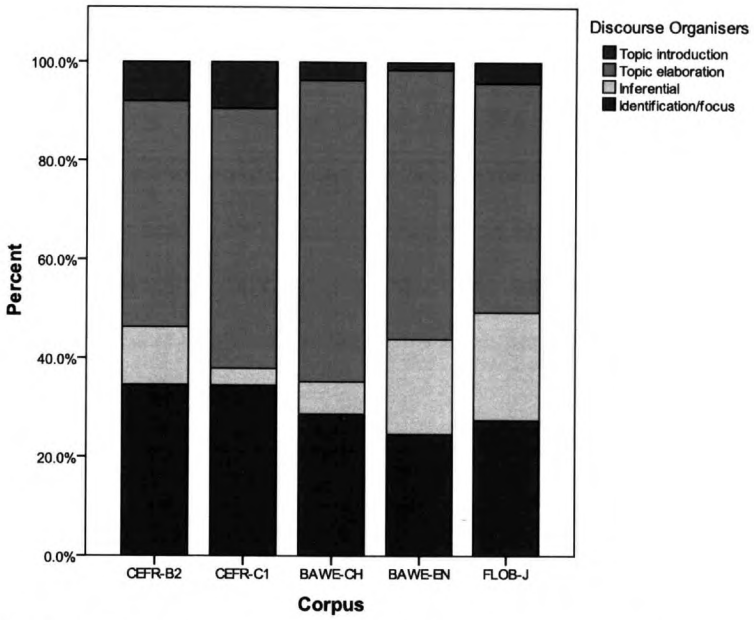


Figure 7-10 Distribution of discourse organising subcategories (tokens-percentage)

We now look more closely into the inferential bundles, which are used to make an inference. All the inferential bundles are grouped on the basis of the core words, the words making the most contribution to the overall meaning in lexical bundles, and the results are presented in Table 7-4. Three inferential bundles are exclusively found in expert academic writing FLOB-J only: *in the light of*, *in the sense that* and *in view of the*. These three bundles evoke a certain tone of formality which seems to be typical of formal/academic writing. On the other hand, the inferential bundles incorporated with the core word *result(s)* (e.g. *as a result of*) are widely used by the writers across different proficiencies, and they appear to be compatible with either academic writing or EAP-like writing. It is also found that proficient learner writing in CEFR-C1 and BAWE-CH reveals very little reliance on these inferential expressions. It is likely that the advanced learners resorted to other markers of inference composed of single words (e.g. *therefore*, *so*) or shorter clusters (e.g. *because of*), and the

collocated word combinations fail to reach the frequency threshold. In comparison, the British student writing represented in BAWE-EN appears to enormously favour expressions incorporating *due to*, resulting in four variations of such bundles. Remember the excessive preference of *use/used* found in the recurrent word combinations in BAWE-EN (e.g. *through the use of*, *be used in the*), which accounts for as much as 11.5% of the total bundles in British student writing (cf. Chapter 5). This elusive anomaly in overusing certain core words in variant forms of bundles seems to be one unique characteristic in native student writing.

Table 7-4 Inferential bundles in Modular Studies 1 and 2

Epistemic bundles	Corpus (Word count) / Freq	Modular Study 2		Modular Study 1		
		CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
because	because they are not	4				
	because it is not				4	
	the main reason is	5				
due to	due to the fact+(that)				8	
	this is due to+(the)			4	5	
	this may be due to				4	
	to a lack of[41]				5	
result	the result of the	4				
	+the result of this	5				
	+the results of the					4
	as the result of+	4				
	as a result of		4	12	17	9
	and as a result				7	
(Expert)	in the light of					6
	in the sense that					8
	in view of the					4

[41] The collocated words preceding the bundle '*to a lack of*' are *due (to)*, *down (to)*, *in response (to)*, *lead (to)*, which are more or less used to express inference. This unique bundle has been discussed in Section 4.3, which deals with functional categorisation.

7.1.3 Good Indicators for Writing Competency

In Chapters 5 and 6, chi-square tests were carried out between the corpora compared to attest whether there is a significant difference in structural or functional distribution of lexical bundles. Then the chi-square standardised residuals were calculated to examine which categories made a major contribution to the difference.

The overall results of chi-square standardised residuals in structural comparisons in Modular Studies 1 and 2 are presented in Table 7-5 and Table 7-6. The cells with an absolute value of standardised residual *R* greater than 1.96, which suggest the major contributors to rejecting the null hypothesis, are indicated with shading. As can be seen, the extent of variation is always more pronounced in the token distribution than the type distribution. In terms of structural distribution in two modular Studies, however, there does not appear to be a consistent pattern. In general, NP-based and VP-based bundles make more contribution to the significant difference among the academic writing in Modular Study 1 whereas the category of PP-based bundles is the primary contributor to the significant difference between two CEFR learner groups of writing in Modular Study 2.

Table 7-5 Overall chi-square standardised residuals in structural distribution (types)

Type		***NP-based***	***PP-based***	***VP-based***
Study 1	BAWE-CH	-1.1	0.1	0.7
p=0.005, Significance	BAWE-EN	-1.2	-0.7	1.5
	FLOB-J	**2.1**	0.7	**-2.0**
Study 2	CEFR-B2	0.2	-1.1	0.5
p=0.12, No significance	CEFR-C1	-0.3	1.6	-0.7

Table 7-6 Overall chi-square standardised residuals in structural distribution (tokens)

Token		*NP-based*	*PP-based*	*VP-based*
Study 1	BAWE-CH	**-3.4**	1.5	1.0
p<0.0005, Significance	BAWE-EN	**-2.2**	**-3.0**	**4.5**
	FLOB-J	**5.0**	1.7	**-5.1**
Study 2	CEFR-B2	0.8	**-2.7**	1.3
p<0.0005, Significance	CEFR-C1	-1.1	**3.6**	-1.8

If we just focus on the positive and negative values of *R*, which are indicative of higher counts than expected (positive value) and lower counts than expected (negative value) regardless of significant difference, one general pattern shared by these two studies is that the more proficient writers as a whole used fewer VP-based bundles and relied on PP-based bundles more (i.e. FLOB-J and CEFR-C1). As to NP-based bundles, the pattern is not as clear as the other two categories.

The overall chi-square standardised residuals in functional comparisons in Modular Studies 1 and 2 are displayed in Table 7-7 and Table 7-8. Similarly, we see the pattern observed in the token distribution (Table 7-8) demonstrates the variation with a greater extent when compared to the type distribution (Table 7-7). In terms of major contributors to the significant difference, the categories of referential expressions and discourse organisers are overall the ones that can better distinguish expert writing and novice/learner writing in Modular Study 1. When frequency is taken into account, then stance bundles in two groups of student writing also made a substantial contribution to rejecting the null hypothesis. As for Modular Study 2, there is no significant difference in either type or token distribution, and thus no cells have the absolute value of *R* higher than 1.96.

Table 7-7 Overall chi-square standardised residuals in functional distribution (types)

Type		***Referential expressions***	***Stance bundles***	***Discourse organisers***
Study 1	BAWE-CH	-0.7	-0.7	1.4
p=0.003, Significance	BAWE-EN	-1.5	0.9	1.0
	FLOB-J	**2.1**	-0.3	**-2.2**
Study 2	CEFR-B2	-0.1	-0.3	0.3
p=0.781, No significance	CEFR-C1	0.1	0.4	-0.4

Table 7-8 Overall chi-square standardised residuals in structural distribution (tokens)

Token		***Referential expressions***	***Stance bundles***	***Discourse organisers***
Study 1	BAWE-CH	**-2.7**	**-2.9**	**5.3**
p<0.0005, Significance	BAWE-EN	**-3.7**	**3.2**	**2.0**
	FLOB-J	**5.9**	-0.6	**-6.4**
Study 2	CEFR-B2	0.9	-1.0	-0.3
p=0.07, No significance	CEFR-C1	-1.2	1.3	0.3

The findings in this section as a whole suggest the following conclusion. In structural analysis, the categories of NP-based and VP-based bundles can better distinguish expert and non-expert writing while the category of PP-based bundles can better distinguish CEFR-B2 and CEFR-C1 writing. In functional analysis, referential expressions and discourse organisers are the categories that may be used to distinguish groups of writing in Modular Study 1 although the distinguishing power is not as strong as structural analysis as almost every cell in the token distribution contributed to rejecting the null hypothesis except for the cell representing stance bundles in FLOB-J. As for Modular Study 2, there is no significance difference in either the type or token distribution, and there appears to be no explicit patterns in any tendency towards using more or fewer bundles in certain categories, either.

The parallel of contribution indicated by residuals between structural and functional

categories, at least in Modular Study 1, probably can be attributed to the close interaction between structural and functional categorisation to a large extent as described in Sections 5.5 and 6.6. Combined with the results in relation to bundle structures and functions in Chapters 5 and 6, we know that one feature shared between the two groups of CEFR learner writing and the two groups of university student writing is the dominance of VP-based bundles and discourse organisers. This feature is distinct only in the written samples produced by inexperienced writers such as university students or L2 learners, as native expert academic writing from FLOB-J was found to mainly rely on PP-based and NP-based bundles and referential expressions. It is conspicuous that FLOB-J is the only compilation of mature writing among all the written subcorpora investigated in this project, which distinguishes itself from other samples written by students or learners, native or non-native alike.

Another observation which attracts more attention is the unexpected results from native student writing represented in BAWE-EN. Remember the original assumption was that native student writing would be closer to expert writing in terms of bundle use as English is their mother tongue. Yet the results of chi-square residuals indicate that the quantitative difference contributed by this native student group is sometimes even much greater than that by the L2 student group (e.g. VP-based bundles and referential expressions in the token distribution). This finding looks odd at first sight. The actual use of bundles in each category, nevertheless, also has to be considered. In fact, in the careful examination of each structural and functional category (Sections 5.4.4 and 7.1.2), we have also seen some similarities between native student writing and expert writing. For example, native students used extent/degree modifiers (e.g. *the extent to which*) and also a wider range of hedging devices (e.g. *are more likely to*), both of which are characteristics of native expert writing but not learner writing. It can be concluded that native student writing appears to be similar with L2 student writing not only in the over-reliance on VP-based bundles and discourse organisers

but also in the poverty of NP-based bundles and referential expressions. When it comes to certain aspects of bundle use, i.e. careful control of qualification and certainty, British student writing still shares more similarity with expert writing as opposed to non-native student writing.

As can be seen, chi-square standardised residuals can provide a quick effective overview of comparison, which can throw some light on quantitative difference between groups of writing. Yet such quantitative analysis only involves number of bundles in each structural and functional category but cannot illustrate the actual difference of lexical bundles and the extent to which the bundle use is qualitatively different between groups. For example, under the same subcategory of referential quantifying expressions, CEFR-B2 writing contains 'colloquial bundles' such as *there are a lot of* whereas FLOB-J writing has 'literate bundles' such as *a wide range of*. In other words, such quantitative analysis can only point to a rough tendency of structural distribution, which still needs to be complemented by close scrutiny of bundles and concordances in each corpus. For future research, it might be feasible to distinguish between colloquial bundles and literate bundles before conducting quantitative analysis in order to reflect the actual difference and/or similarity.

7.1.4 Keyness Analysis

As discussed earlier, a keyness analysis can provide a quick overview of the lexical bundles which occur significantly frequently or infrequently in statistical terms when compared with a reference corpus, and it also takes into account of the corpus size. Native expert academic writing FLOB-J is used as the reference corpus for the four student and learner subcorpora as it is considered the standard written English that all novice writers, regardless of L1 or L2, generally intend to achieve in spite of the genre variation between subcorpora. The results turn out to be rather interesting in the sense that the four student and learner corpora have commonly overused and underused certain expressions. The expression *is one of the* is the

only overused bundle in all four non-expert subcorpora, occurring with a very high raw frequency ranging from 9 to 18 (see Table 7-9 below). This bundle, categorised as an identification/focus bundle under discourse organisers, was missed in the previous sections of detailed discussion.

Table 7-9 Key bundles shared in student and learner corpora with FLOB-J as the reference corpus

Key bundles	Freq \ Corpus (Word count)	Modular Study 2		Modular Study 1		Reference
		CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
overuse	*is one of the*	18	14	9	12	--
	for a long time	11	6	--	--	--
underuse	*as a function of*	--	--	--	--	15

A scrutiny of the concordance listings suggests that immature writers, native or non-native alike, appear to use the expression *is one of the* as a kind of hedging device (despite being categorised as a discourse organiser conforming with Biber & Barbieri (2007)) to mitigate the magnitude of the claims they made. The construction incorporated with *is one of the* generally collocates with the superlatives. See the examples below:

- The threat of open source software containing Trojan horses or backdoors ***is one of the*** biggest limiting factors for open source adoption. (BAWE-EN)
- If we can rely on a certain friend during our hard time, it is ***one of the*** finest feelings of human beings. (CEFR-C1).
- The education system is recognized by the public that it ***is one of the*** worst systems in the world. (CEFR-B2)

It is not the case that native expert writers never use the expression *is one of the*. Actually the past tense equivalent *was one of the* is found in FLOB-J, but it occurs only four

times. Compared with FLOB-J, however, *is one of the* is among the most commonly used bundles in the student or learner subcorpora. If ranked on the basis of frequency, it is the most frequently used bundle in CEFR-B2, coming second place in CEFR-C1, and tenth place in both BAWE-CH and BAWE-EN while *was one of the* in FLOB-J barely reaches the frequency threshold of four times.

The other overused bundle, *for a long time*, is found in EAP-like learner writing CEFR-B2 and CEFR-C1 only. As discussed, the expression *for a long time* is a way in which learners intensify their propositions and generally carries a hyperbolic tone which leads to an impression of overstatement found exclusively in learner writing. In addition, the expression *for a long time* sounds rather vague, which is also different from the explicitness and exactness required in academic writing,

- This phenomenon has become more and more serious, and the government has talked about it ***for a long time.*** (CEFR-B2).
- If a person is exposed to a particular language-using environment ***for a long time,*** it gives him opportunity to use the language when communicating with the other people. (CEFR-C1)

As for the only underused bundle *as a function of* in the student or learner corpora, it is understandable why it occurs so frequently in FLOB-J only as a large number of academic texts included in the FLOB-J subcorpus are hard-science based although the texts included in FLOB-J are still considered to be widely distributed across various disciplines (cf. Section 5.1). The words collocated with *as a function of* in FLOB-J are accordingly found be mostly science-related, such as *temperature*, *time*, *size*, or *stand height*. The impact of using FLOB-J as a reference corpus to be compared with students' academic written English will be further addressed in this chapter.

7.1.5 Overall Development & Some Methodological Issues

As discussed in Chapter 2, learner language is generally considered to become more fluent, complex, and accurate as it progresses. In accordance with the rationale behind this thesis, it would also be reasonable to add one more determinant to the construct of L2 development, i.e. the extent of *formulaicity*. It is expected that the number (in terms of types and tokens) of formulaic expressions, frequency-driven as defined in the current study, would increase with proficiency. This assumption appears to be supported by the overall growth of lexical bundles throughout the supposedly least proficient CEFR-B2 learner writing to the most proficient native expert writing FLOB-J. Since the subcorpora investigated in Modular Study 1 are nearly twice as large as the ones in Modular Study 2, it was decided to normalise the number of bundle types and tokens in each subcorpus on the basis of per 100,000 words for the sake of comparability.[42] As can be seen in Figure 7-11, the use of lexical bundles increases steadily across proficiency development with the exception of CEFR-B2 learner writing, and the upward curve stabilises at the two groups of native writing. The strikingly frequent use of lexical bundles in CEFR-B2 writing draws the attention immediately. The recurrent word combinations in CEFR-B2 writing, however, have been found to be fundamentally different from other groups of writing, particularly in stark contrast with native expert writing FLOB-J. The CEFR-B2, which comprises the weakest writing considered, contains an excessive number of colloquial quantifiers (e.g. *a lot of time*, *there are too many*) as well as a number of expressions exclusively used by L2 writers (e.g. *for a long time*). It could be hypothesised that something interesting is happening to learners as they progress from stage B2 to C1. They abandon a range of 'inappropriate' bundles (which for academic writing would be considered too vague, generalising or colloquial), which results in their overall use of types

[42] Normalising the number of bundle types with corpus size could be disputable, but this practice will be further justified later in this section.

and tokens of bundles to be considerably depleted. Then, over time, they incorporate a new set of 'academic' or 'literate' bundles into their writing, resulting in a gradual increase in both types and tokens.

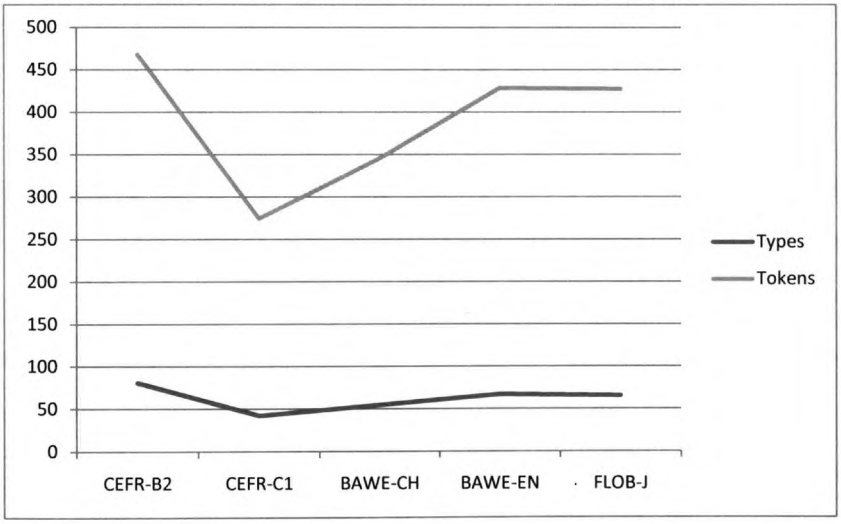


Figure 7-11 Overall number of bundle types and tokens per 100,000 words

Another noticeable phenomenon is that the diversity and frequency of lexical bundles in British student writing appears to be virtually identical to expert writing. Yet, again, remember that British student writing shows an unusually inordinate reliance on expressions with *use/used* as the head word (e.g. *for the use of*, *will be used to*), which constitutes approximately 11.5% of its bundle types (cf. the discussion in Section 5.3.4.1). Given the idiosyncrasies observed in CEFR-B2 writing and BAWE-EN writing, it can be concluded that on the whole the diversity and frequency of lexical bundles tend to increase with writing proficiency development. If CEFR-B2 is not included in the curve, this graph also seems to suggest that initially the use of phraseological expressions becomes more diverse and frequent steadily throughout language development; then the upward curve become much

gentler after reaching a critical maturity point.

In fact, such a pattern of growing use of formulaic expressions discovered by this present study contrasts sharply with a couple of existing bundle studies which report that non-native writing entails more recurrent word sequences than native writing. With the aim to differentiate academic genres rather than L1 and L2 writing by means of lexical bundles, Hyland (2008a) compared the published research papers produced by native academics and the masters theses and doctoral dissertations written by L1 Cantonese students of L2 English in Hong Kong. He found that the masters students used the greatest number of lexical bundles while the academics used the least. Hyland speculated that less confident students might make more effort to construct their texts by employing more formulaic expressions to display their competence. A number of methodological issues in Hyland's study, nevertheless, could undermine this conclusion. For one, the retrieved word sequences do not appear to be scrutinised for overlaps or context dependence. We thus see some possible overlapping clusters such as *it should be noted/should be noted that* and *at the beginning of/the beginning of the* and also context-dependent clusters such as *of the Hong Kong* and *in Hong Kong and*, which were called 'research-oriented' clusters by Hyland. For another, it is also rather perplexing to find word combinations like *the other hand the* and *in other words the* in Hyland's data. A comma is quite likely to have been placed prior to the article *the*, thereby with the possible original forms as *(on) the other hand, the* and *in other words, the*. Accordingly these two discontinuous clusters interrupted by a comma should not be included as a target word combination, since a string of words across the boundary of punctuation presumably falls out of the scope for investigation. Finally, without a thorough examination of individual bundles in each group, it cannot be certain whether the number of bundles in postgraduate writing corpora is inflated as a result of various forms originating from a base form (such as the excessive reliance of *use/used* in twelve bundles observed in BAWE-EN,

e.g. *can be used to*, *also be used to*). In Hyland's data, for instance, several pairs of bundles found in postgraduate writing, such as *it was found that/it is found that* or *in terms of the/in terms of their*, actually originate from the same base form (i.e. *it* BE *found* and *in terms of* NOUN in the above cases), which could substantially inflate the number of bundles. On the other hand, such variation of a bundle form is observed less often in the corpus of expert writing. These details appear minor yet could be crucial if considered together, and could therefore alter the number of retrieved bundles in Hyland's study and hence impact on the result.

De Cock (2000) also took advantage of the same frequency-driven automated approach in terms of comparing L2 data with L1 data in writing as well as in speech. She investigated the repetitive chunks from an overall perspective of looking at all the continuous word combinations ranging from two-word to five-word lengths without further refinement or categorisation. The patterns emerging from De Cock's data reveal that learners tend to use more highly repetitive phrasal chunks in both speech and writing than native speakers. As De Cock herself pointed out, however, 'a structural classification and a thorough functional investigation of the combinations in context is required before they can be labelled as such [prefabricated expressions]' (De Cock, 2000, p. 59). It is believed that the approach adopted in this thesis has provided a feasible remedy for the procedural problems arising in De Cock (2000) and Hyland (2008a) in the sense that the lexical bundles investigated here could better reflect the genuine building blocks in constructing a text as opposed to being mixed with some repetitive proper nouns or overlaps. The repertoires of native and learner bundles have been not only manually scrutinised for overlaps and context dependence but also categorised through a sound structural and functional framework. Consequently, the overall growth of formulaicity with the advance of writing competency, which is the case for both the diversity of lexical bundles (types) and the frequency (tokens), should be more plausible than reported

in the literature.

Certainly, direct comparison between studies has to be tackled with extreme caution as various operational definitions were adopted in different studies, which would have a great impact upon the number of recurrent word combinations retrieved. As explained in Chapter 3, three competing factors can affect the number of bundle types retrieved to a large extent, i.e. cut-off frequency, dispersion requirement, and corpus size. The repeated experiments in this thesis reveal that as a whole, larger corpora would generate fewer recurrent word combinations with the same cut-off normalised frequency when compared with smaller corpora because large corpora would accordingly end up with a higher converted raw frequency. In comparison, a fixed dispersion requirement, say word sequences occurring in at least three texts or five texts, usually does not affect the retrieval of number of word combinations in large corpora as much as in small corpora because lexical bundles extracted from large corpora with the same cut-off frequency generally cover a much wider range of different texts (cf. Biber et al, 2004, 2007). Yet it is not clear how using an unusual dispersion requirement as in Hyland's study (2008a), i.e. requiring a bundle to occur in 10% of texts, would impact on the number of recurrent word combinations. In theory, it is ideal to compare corpora of exactly the same size composing of the same number of texts, but this is virtually unachievable in reality. It is thus recommended to use proportional data for intra-study quantitative comparisons as opposed to raw frequencies. For cross-study comparisons, we have to bear these limitations in mind, particularly when the student groups in my studies are not identical to Hyland (2008a).

It also has to be acknowledged that normalising the number of different lexical bundles (i.e. bundle types) with a standardised corpus size as presented in Figure 7-11 might be disputable (Biber, 2006; Biber & Barbieri, 2007). The principle underlying the numbers of bundle types and tokens is similar to that of single words. The range of all different words or

phrasal units is supposedly finite in our mental lexicon whereas the number of their occurrences is infinite, subject to text length. In other words, the number of tokens is linear with corpus size while the number of types is not. With respect to lexical bundles, this issue is even more compounded with the addition of two more variables: frequency and dispersion thresholds. The complex interaction between bundle types, corpus size, and frequency and dispersion requirements has been explored in Section 3.5.1. For this section, the purpose of normalisation is simply to better present the results of comparison when taking into account unequal corpus sizes with the same cut-off raw frequency (thereby different converted cut-off normalised frequency). As can be seen in Table 7-10, the set frequency threshold of four times in determining lexical bundles would result in a lower normalised figure per million words for a much larger corpus like FLOB-J, which could raise some doubts as to whether this is why FLOB-J generates the highest number of lexical bundles. To lessen the impact from corpus size, it was thus decided to normalise not only bundle tokens but also bundle types. In Table 7-11, it can be seen that the overall trend before and after normalisation is still largely identical in terms of bundle types, simply with a slight adjustment of magnitude. On the part of bundle tokens, the aspect which hinges on corpus size, we can now see a picture of a more sensible linear relationship after normalisation, as discussed earlier and visually presented in Figure 7-11.

Table 7-10 Constituents of corpora and cut-off frequency for determining lexical bundles

Corpus	Corpus size (words)	No. of texts	Average text length (words)	Raw cut-off frequency	Normalised cut-off frequency (per million words)
CEFR-B2	87,970	239	368	4	45.5
CEFR-C1	87,828	157	559	4	45.5
BAWE-CH	146,872	53	2,711	4	24.3
BAWE-EN	155,781	60	2,596	4	25.7
FLOB-J	164,742	80	2,059	4	27.2

Table 7-11 Bundle types and tokens before and after normalisation (per 100,000 words)

Corpus	Types		Tokens	
	Raw	Standardised	Raw	Standardised
CEFR-B2	70	80	411	468
CEFR-C1	37	42	241	274
BAWE-CH	80	54	511	345
BAWE-EN	104	67	667	428
FLOB-J	108	66	704	427

The last issue to be addressed in this section is the proportion of formulaicity in language. As indicated in Table 7-12, in the five groups of writing investigated, one lexical bundle occurs per 214-364 words. These recurrent word combinations, mostly four-word units with a few exceptions consisting of five or six words, were multiplied by four to generate approximate total word counts so that the proportion of words consisting of bundles in each subcorpus could be calculated. The result shows that the lexical bundles constitute about 1-2% of the total running words in each subcorpus. Take the university student essays extracted from the BAWE corpus for example. Accordingly, a typical 2000-word student essay would entail approximately seven to nine four-word lexical bundles, functioning as building blocks to construct the academic discourse. This looks like a tiny proportion compared with what has been reported in the literature (e.g. 52% in Erman & Warren, 2000), but bear in mind that these are highly recurrent word sequences, which must occur at least four times (about 25-45 times per million words) in three texts or more. If the cut-off requirement is set at a lower threshold, say 10 times per million words such as in Biber et al. (1999), more word combinations would be retrieved than those in the current repertoire. Meanwhile, it is probably also not fair to calculate the proportion of multi-word units on the basis of running words in a corpus as the semantic units are generally not established on individual words. Yet before a more sensible calculation, if any, can be devised, the current

way of calculation seems to be the only solution.

Table 7-12 Spread of lexical bundles (tokens)

Corpus	1 lexical bundle token per ? words	% of total running words
CEFR-B2	214	1.9%
CEFR-C1	364	1.1%
BAWE-CH	287	1.4%
BAWE-EN	234	1.7%
FLOB-J	234	1.7%

It is also very likely that the intensity of formulaic chunks disclosed by such an automated frequency-driven approach has been under-represented as only the exactly identical forms would be retrieved. In other words, any variations of a base form of formulaic units would be missed if the accumulative frequency does not reach the threshold. For example, if an adverb such as *very* was inserted into the base form of a lexical bundle *it is (very) difficult to*, this case would not be picked up by the automated procedure. Other possible varied forms include lexico-grammatical variations such as verb tense (e.g. *it is suggested that/it has been suggested that*) or the singular/plural noun distinction (e.g. *than that of the/than those of the*). On the other hand, the same observation also raises another caveat that certain types of lexical bundles might have been over-represented in the quantitative analysis for bundle types as one base form of a lexical bundle can have a number of variations. Several typical examples found in the data include *it can be argued* and *it could be argued*, *the way in which* and *the ways in which*, *has a lot of* and *have a lot of*, and *is likely to be*, *are likely to be*, and *are more likely to*. It has to be acknowledged that the majority of variations of a base form are still treated as separate lexical bundles despite their structural

and semantic origins.[43] More often than not, it is not easy to determine whether two or more varied forms should be combined or not. For example, given that *it can be argued* and *it could be argued* express different degree of epistemic modality, one might argue that both forms should remain in the data. For further research, perhaps one way to deal with this issue is to scrutinise the corpus data to search for all the possible variant forms and then devise a systematic categorisation scheme so as to more accurately present a quantitative analysis.

Another problem in relation to under-representation and over-representation is learner errors. As the Longman Learners' Corpus is not error-tagged, it is not clear to what extent the spelling or grammatical errors could impact on the results of retrieval. In effect, the only erroneous form that came to light as a retrieved word combination is **a lot of problem* (cf. Section 6.4.4.1). In the Longman data, other learner bundles associated with errors are the quantifiers with mismatch nouns (e.g. *a large amount of* collocated with a countable noun *different culture*) or the structure '*there is/are* + NP' structure collocated with erroneous verb forms (e.g. **there are so many* collocated with a noun plus a finite verb *prostitutes exists*) (cf. Section 6.5.4.1). These error-prone bundles found in learner writing, however, generally do not entail visible errors within the bundles. In the case of the BAWE corpus, the only problematic learner bundle found is *?in the recent years* (cf. Section 5.4.4.1) with a possibly redundant article *the*. As the BAWE corpus targets at student writing within the British higher education as opposed to L2 writing, learner errors are not considered.[44] Therefore, what remains unclear is the errors which do not form part of a lexical bundle or collocated with

---

[43] The only exception is the two retrieved learner bundles **a lot of problem* and *a lot of problems*. The former is the erroneous form while the latter is the correct form. As **a lot of problem* was generated as the result of a learner error, it has been combined with the correct form so as to avoid data inflation.

[44] It could be argued that errors would not be an issue for the BAWE data because the assessed essays collected are considered to be proficient student writing regardless of L1. Certainly, L2 learners would still make errors even if they have reached a very advanced level, and probably these errors would be more difficult to be spotted (e.g. collocations or stylistic infelicities). This could be another direction for future research.

one, thereby being overlooked. It is speculated that learner errors are more likely to lead to recurrent word combinations being under-represented for the deviant forms could not possibly be detected automatically. Additionally, the irregularity and unpredictability of learner errors makes it impossible to make the same contribution as the correct forms do.

As has been discussed, the lexical bundles investigated are highly recurrent and have been carefully filtered to remove any 'disqualified' word sequences, which could explain the relatively low proportion of lexical bundles in the total running words to some extent. Meanwhile, in an attempt to uncover why this proportion statistic is so drastically different in various studies (see Table 7-13), it is also worth comparing how other researchers generally calculated the proportion of formulaicity in their data.

Table 7-13 Proportion of formulaicity reported in the literature

Register	Study	Proportion of Formulaicity
Speech	Altenberg (1998)	80%
	Erman & Warren (2000)	59%
		52%
Writing	Li & Schmitt (2009)	2-3%
	Modular Studies 1 & 2 in this thesis	1-2%
Mixed	Moon (1998a)	4-5%

One of the most widely cited proportions, and the highest one to date, is 80% in spoken English reported by Altenberg (1998), who investigated 'any continuous strings of words occurring more than once in identical form' (*ibid*, p.101). This argument inevitably does not take into account both overlaps and context dependence prevalent in automatically retrieved recurrent phraseology, which makes this high figure very misleading. Taking a set

of broadly defined criteria which include lexical, pragmatic, grammatical, and reducible prefabs, Erman and Warren (2000) also reported very high figures for the corpora investigated, 59% for speech and 52% for writing. The rest of studies which reported the proportion of formulaicity largely adopted a top-down approach instead, which started with native speakers' judgment or a pre-existing list of fixed expressions to check against the data, and the proportion of formulaicity calculated in this manner turns out to be closer to the 1-2% reported in this thesis. For example, checking against a list of 6,776 formulaic sequences, Moon (1998a) estimated that only around 4% to 5% of the words in her corpus were parts of fixed expressions. In a longitudinal case study which investigated one L2 learner's acquisition of formulaic language in academic writing, Li & Schmitt (2009) reported a frequency of one lexical phrase token per 35-49 words or 2-3% of the total running words in their sole target learner's written output. To ensure the greatest percentage of lexical phrases to be covered, their formulaic sequences were very leniently identified. As long as one of the three expert judges considered the sequences to be formulaic, the sequences were added to the repertoire. However, the figure 2-3% reported by Li & Schmitt (2009) was found to be rather ambiguous. They seemed to have simply divided the occurrences of lexical phrases with the total running words while the units for comparison, phrases vs. words, were not comparable. Moreover, the claim that their percentages were very similar in comparison with the proportion 5% on the part of L2 writing in Howarth's study (1998b) is also perplexing. The phraseological units discussed in Howarth (1998b) are various types of 'verb + direct object' collocations, no occurrences of which have ever been mentioned in this paper. The comparability of units (types vs. tokens and verb + object vs. lexical phrases) between Howarth's study (1998b) and Li & Schmitt's study (2009) is thus very questionable. Furthermore, the native data investigated by Howarth has an even lower proportion for collocation types 2% when compared with learner writing. Li and Schmitt (2009)'s claim that

their proportion is in line with Howarth's learner data, thus the extent of L2 formulaicity being much lower than that in native writing, is hence also invalid and can be very misleading. The thorough examination of the above phraseology studies raise a serious issue as to how the researchers calculated or estimated the proportion of formulaicity in their data. The comparability of results involves two kinds of units for comparison, i.e. phraseological units and types/tokens, which have to be carefully considered before a conclusion can be reached.

To sum up, the results from the current study still indicate an overall progressive relationship between writing development and diversity and intensity of the lexical bundles. In comparison with other studies which also compared L1 and L2 in terms of the use of recurrent word combinations, various aspects of this frequency-driven approach were revisited. It was found that a sensible comparison has to be established upon carefully refined data in conjunction with thorough examination of individual word combinations rather than purely quantitative analysis. A further comparison with other phraseology studies which reported the proportion of formulaicity also suggests that researchers have to be extremely cautious with what units of formulaic sequences are to be compared when referring to the proportion reported in the literature. It has to be made clear whether types or tokens were used and how the formulaic sequences were calculated in proportion to the corpus as corpus size is generally represented with individual words while formulaic sequences vary in terms of length (two words or more) and occurrences.

7.2 Distinctive Features across Proficiencies

The results of comparisons between two Modular Studies have been presented and discussed earlier. In this section, we will see how the above findings correspond to the related research on second language writing and English for Academic English (EAP) and how they can be operationalised in more accessible terms for the ELT community.

7.2.1 Towards a More Complex Language

Recall that the general assumption underlying second language development discussed in Chapter 2 is that the more proficient the learners, the more complex language they would produce. This assumption also appears to hold true in the language represented by lexical bundles, even though the data analysed here are simply a set of frequency-driven word combinations as opposed to full texts. The distribution of structural subcategories of lexical bundles used in the five groups of writers is presented in Figure 7-12 (for the structural subcategories, see Section 4.2). As can be seen, the most proficient writers in FLOB-J made use of the widest range of various grammatical structures while the range of structures shown in the word combinations is less diverse in the four less proficient groups (the darker colours being indicative of better proficiency). The structures that one or more non-expert groups of writing did not make use of in their repertoires of lexical bundles are noun phrase fragments (e.g. *the fact that this*), adverbial clauses (e.g. *as we shall see*), *that*-clauses (e.g. *that there is a*), and verb phrase fragments (e.g. *has a number of*).

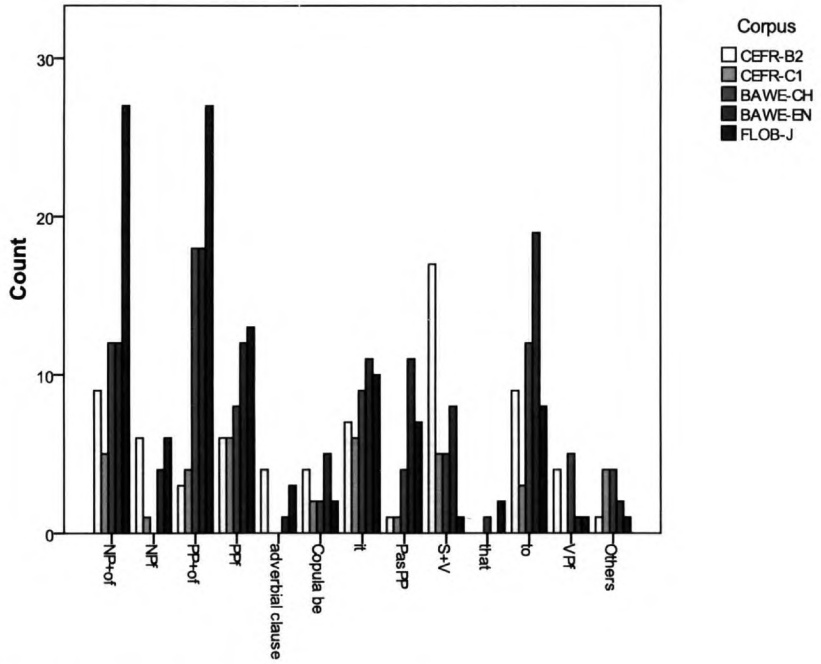


Figure 7-12 Distribution of structural subcategories (types)

We have seen that the more proficient writers, by and large, used more noun-based phrases, preposition-based phrases, and a wider range of structures in verb-based phrases despite the relatively smaller percentage of verb-based bundles (cf. Section 7.1.1). The effect of intensive normalisation in proficient writing resulting from a large number of nominal and prepositional phrases, particularly the ones with the *-of* structure, is a style packed with more information. Interestingly, complexity has been regarded as one fundamental feature that distinguishes speech and writing.[45] Written language is generally considered to be more

[45] The notion of *complexity*, however, was interpreted differently by different researchers. For example, Halliday considered the complexity of written language to be 'static and dense' while that of the spoken language is 'dynamic and intricate' (1989, p.87). He particularly stressed that speech is no less structured than writing in terms of grammatical intricacy, but he also pointed out that writing is characterised with highly information-packed and lexically dense passages.

information-intensive and complex than spoken language (Biber, 1988; Biber et al., 1999; Halliday, 1989). From the perspective of second language development, this seems to suggest that learner writing at lower levels tend to be more speech-like while more proficient learner writing is closer to the norm of written language recognised in the native community. In the following sections, we will see that this assumption is well supported by more evidence.

7.2.2 Cliché & Verbosity vs. Concise Language

In the qualitative analysis in Chapters 5 and 6, it has been found that learners tend to overuse certain fixed expressions, particularly the highly frequent ones, to the extent that they might begin to sound unnatural and cliché-ridden. Table 7-14 shows a selection of lexical bundles that appear to be 'abused' by the learner writers. The determination of clichéd bundles is somewhat arbitrary. The first stage was based on an examination of bundle repertoires in the three groups of learner writing, in which any bundles with a tone of cliché and banality were flagged as candidates. Then the frequencies of these candidates were crosschecked against those in the two native groups. The bundle candidates with a frequency higher than any native group by at least four occurrences were defined as clichéd expressions. As the frequency of four is one of the thresholds for defining a lexical bundle in this thesis, and the two CEFR learner subcorpora are in fact much smaller than the two native corpora, this measure is considered to be effective in distinguishing overused expressions from other bundles. As can be seen in Table 7-14, for those widely known formulaic expressions shared by native writing, learners used them far more frequently (e.g. *on the other hand*, *at the same time*). Some of the expressions favoured by the learners sound somewhat stilted (e.g. *last but not least*), which were probably learnt by rote from their ESL learning materials or classroom instruction.

Table 7-14 Clichéd bundles

Bundle / Corpus (word count)	Freq	Modular Study 2		Modular Study 1		
		CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
on the other hand		18	28	36	4	19
at the same time		17	14	24	5	10
as a matter of (fact)		4	5	--	--	--
all over the world		5	5	6	--	--
for a long time		11	6	--	--	--
in the long run		--	--	13	--	--
(played) an important role in/ a very important role		4	--	5	--	--
last but not least		--	--	5	--	--

Another feature that might be relevant to the overused expressions in learner writing is the higher percentage of discourse organisers when compared with native expert writing (see the discussion in Section 7.1.1). It appears that learners rely a lot more on explicit discourse connectors to structure and elaborate their text than native expert writers. By contrast, the mastery of utilising as few words as possible aiming for a concise style may be regarded as an expected skill for academics as journal or book publication usually requires the authors to adhere to word limits. Consequently, in the context of academia, short and clear expressions are generally preferable to long-winded word sequences. In line with this finding, Kennedy and Thorp (2007) also discovered that IELTS writers at lower levels used lexico-grammatical markers (e.g. *however*) more frequently than advanced writers. They found proficient L2 writers actually relied on other means (which is still unclear and thus requires more research) to structure their arguments, appearing to be nearer to native speaker use in this respect.

From the observations of use of lexical bundles, verbosity is another impression formed about learner writing, namely that learners often use more words than required to convey an idea. Part of the impression originates from a few word combinations employed by

novice writers, native and non-native alike (cf. Section 5.4.4.3), which indicates that student writers tend to repeat the same point with different words (e.g. with the use of *that is to say*) or use a wordy expression when a shorter form may be available for the same function (e.g. *(due)+to the fact that* instead of *because*, *as a matter of fact* instead of *in fact*). One characteristic also found in learner writing at the lower proficiency is the existential '*there is/are* + NP' pattern, the pronoun use of '*it*', and the 'copula BE' structure, which are commonly seen in CEFR-B2 bundles (e.g. *there are still some, it is a good, is very important to*). As described in Section 6.5.4.1, many occurrences of the '*there is/are* + NP' pattern bundles followed by a finite verb were actually the consequences of learner errors at the clause level. The existential structure '*there is/are* + NP' is used to stress the notion of existence (Quirk & Greenbaum, 1973, p. 418). Yet the immoderate use of this structure as well as superfluous appearance of copula BE in writing gives rise to a style that appears both simplistic and verbose. In fact, in one academic writer's handbook, writers are advised against using sentences beginning with '*there is/are*' and '*it is*' structures, and BE and HAVE also should be avoided being used as main verbs if alternative stronger verbs exist (Rosen, 2008, pp. 425-428).

In conclusion, over-reliance on overt discourse organisers and repeatedly using a limited range of familiar formulaic sequences is perhaps one of the reasons why non-native writing can still sound unnatural or awkward even at a very advanced level. In comparison, native expert writers are able to not only employ a wider range of expressions in a more modest manner but also convey their ideas with as few words as possible. The implications for ESL/EFL learning are twofold. In addition to acquiring new formulae to replace those overused ones, ESL/EFL learners would also need to know how to use the formulaic expressions more moderately. Sometimes perhaps semantic coherence alone is sufficient for any effect that overt discourse markers can achieve.

7.2.3 Colloquialism vs. Formality

As discussed earlier, in terms of structural complexity, less competent L2 writing appears to be more speech-like and more proficient L2 writing tends to be closer to the written norm recognised by native speakers. We have seen that a number of lexical bundles used only by learner writers suggest a colloquial tone, which is undesirable in any context of formal writing including academic prose. Colloquialism in learner writing, particularly at the CEFR-B2 level, is indicated by features such as vagueness (e.g. *some people think that*, *a lot of time*) and personal involvement (e.g. *I am going to*, *I think it is*, (*from*) *my point of view*). Colloquialism is also suggested by the undue use of emphatic adverbs such as *totally* or *very* (e.g. *is totally different from*, *a very important role*, *as we all know*). In comparison with CEFR-B2 writing, the extent of colloquialism decreases sharply in both CEFR-C1 and BAWE-CH writing, which can be evidenced by the sparse distribution of first person pronouns and emphatic adverbs in the lexical bundles of these two groups. Other studies have also shown that learners tend to overuse certain groups of words and expressions more typical of native speech in their writing, e.g. person pronouns and short adverbs (Granger & Rayson, 1998), *I think* (Granger, 1998), and *from my point of view* (Gilquin & Paquot, 2007). Overt writer visibility is particularly noticeable in the least proficient CEFR-B2 writing with the most lexical bundles with the use of first person pronouns. The relationship between the use of first person pronouns and writing quality will be further explored later in the discussion of authorial voice.

Turning to formality demonstrated in academic writing, one of the prominent attributes is the extremely dense distribution of noun and preposition phrases, particularly those embedded with *of* (e.g. *in the presence of*, *the creation of the*). As discussed earlier, the highly intense nominalisation in the written genre has been extensively described in many studies which aimed to identify distinctions between speech and writing (e.g. Biber, 2006; Biber, et al., 1999; Halliday, 1989). Summarised from the findings in Chapters 5 and 6, a

number of groups of lexical bundles are also regarded as markers for formal or academic writing. The first group is extent/manner modifiers (e.g. *the degree to which*, *in so far as*, *the way(s) in which*), which were not used by the three groups of learner writers. The second group is inferential markers which are formal in tone, (e.g. *in the light of*, *in view of the*, *in the sense that*), none of which appear in the lexical bundles in any of the non-expert writing groups. The final group is a wide range of hedging expressions (e.g. *seems to have been*, *are more likely to*), which learners have not shown a sign of mastery even when they have reached the tertiary level as evidenced in BAWE-CH writing.

L2 learners have been found to lack register-awareness (Gilquin & Paquot, 2007), which results in the above speech-like features in learner writing. The findings from the current study suggest that it is important not only to introduce a wide range of appropriate markers which signify a formal written genre into the ESL/EFL curriculum but also to raise the speech/writing distinction wherever is possible in the ESL/EFL curriculum.

7.2.4 Overgeneralisation vs. Cautious Language

Degree of assertiveness is carefully controlled in native academic writing, as can be seen in a diversity of epistemic bundles used in FLOB-J for a scale of probability, starting with the least certainty (e.g. *is more likely to*), to neutral proposition (e.g. *it has been suggested*), to the strongest commitment (e.g. *to the fact that*, *there is evidence that*). This sophisticated skill of being precise yet noncommittal in native academic writing can also be observed from those bundles which are concerned with measurement (e.g. *per cent of the*, *the size of the*, *the magnitude of the*) or the extent modifiers mentioned earlier (e.g. *the degree to which*, *to a large extent*). It is also discovered that native academics were capable of using various measures to qualify their claims such as incorporating modal verbs or hedging nouns in the expressions (see Sections 5.4.4.2 and 7.1.2.2.). The importance of cautious language in the academic context, especially the use and maintenance of hedges, has been addressed by a

large number of studies (e.g. Crompton, 1997; Poos & Simpson, 2002; Skelton, 1988). By contrast, Chinese students in Hong Kong were found to be rather weak in this pragmatic aspect (e.g. Flowerdew, 2000; Hyland & Milton, 1997). Learners' ignorance about cautious language could be ascribed to the failure of addressing this empirical usage in EAP learning materials (Hyland, 1994). In the learner corpora investigated in this thesis, apparently L1 Chinese learners of L2 English have shown some control of this specific feature in the academic discourse, particularly in BAWE-CH writing (e.g. the use of *it is believed that*, *are more likely to*), but they have not demonstrated it as diversely and robustly as native writers did. Interestingly, through a keyness analysis, the expression *is one of the* was found to be markedly frequent in all the four student and learner groups of writing as a mitigating device whereas native expert writers did not use it as frequently.

In stark contrast to L2 writers' underuse of hedging devices is the tendency to be categorical and overgeneralising. In addition to the hyperbolic tone characterised with various quantifiers in CEFR-B2 writing (e.g. *a lot of people*, *there are so many*, see Section 7.1.2.1), other bundles with an overstating tone are listed in Table 7-15. It is interesting to note that the tendency of overgeneralisation appears to attenuate with language development in terms of both range and intensity.

Table 7-15 A selection of overgeneralising bundles

Corpus (word count) / Bundle Freq	Modular Study 2		Modular Study 1		
	CEFR-B2 (87,970)	CEFR-C1 (87,828)	BAWE-CH (146,872)	BAWE-EN (155,781)	FLOB-J (164,742)
all of them are	4	--	--	--	--
as we all know	4	--	--	--	--
become more and more	6				
the most important thing (is)	7	--	--	--	--
is totally different from	4	--	--	--	--
for a long time	11	6	--	--	--
all over the world	5	5	6	--	--

As Ringbom (1998) pointed out, even at the advanced level, learner language is in some respects more, in others less, *vague* than native speaker language, although he set off from a vocabulary-based perspective rather than a phraseological one. Investigating learner writing development with IELTS candidate scripts across band scores, Kennedy and Thorp (2007) also concluded that L2 learners at lower proficiency levels tend to express their opinions in a more categorical manner and their writing is modified less by hedging. The finding here, therefore, reinforces this distinctive aspect of L2 writing from a phraseological viewpoint. The tendency of being less hedged with an overstated tone seems to be universal for learners from different L1 backgrounds as the studies discussed above are not exclusive to L1 Chinese learners of L2 English only.[46] Moreover, it appears that these features could change with proficiency development as evidenced by Kennedy and Thorp (2007) and the present study. Learner writing generally improves as proficiency progresses, most likely by getting closer to the norm shared in native expert writing and with a better control of cautious language. Certainly, for general argumentative and expository writing, the status of cautious

[46] In contrast, an L1 Chinese specific feature seems to be the '*there is/are* + NP' pattern miscollocated with verbs and possibly also the overuse of this pattern in CEFR-B2 writing, as the consequence of mother tongue Chinese.

language may not be as essential as in academic writing. The tendency of overgeneralisation and overstatement revealed from learner writing, nonetheless, appears to be one of the features that makes L2 writing sound foreigner-like.

7.2.5 Shift of Authorial Voice

In Section 7.1.2.2, we have seen that there is a shift of authorial voice embedded in the lexical bundles, from the first person singular pronoun *I* to the first person plural pronoun *we* across writing proficiency (and also possibly across text genres). It has to be emphasised that native expert writing used the collective pronoun *we* in the lexical bundles in a very different fashion compared with learner writing. The native academics utilised personal expressions in adverbial clauses (i.e. *as we have seen* and *as we shall see*) for deictic purposes, directing readers' attention to the flow of the text. On the other hand, the learners used the collective pronoun *we* to identify the focus in the forthcoming proposition (e.g. *as we all know*, *we can see that*, *we can say that*). The impersonal pronoun *it* was also frequently seen in the lexical bundles within the five groups of writing; yet, again, writers across proficiency levels demonstrate different ways of employing it. In the extraposition '*it* pattern', adjectives as well as passives are common in expert writing (e.g. *it is clear that*, *it can be seen that*), while CEFR-B2 writers appear to use this impersonal pronoun more often to refer to some entity mentioned earlier (e.g. *it is a very*, *it is a good*).

Learners have been found to use more personal pronouns in argumentative writing than native speakers (Granger & Rayson, 1998; Petch-Tyson, 1998). Interestingly, Hyland (2002) reported a contradictory finding in L2 academic writing. He compared the undergraduate theses collected from Hong Kong with a corpus of published research articles and also conducted interviews with the students and their supervisors. Hyland concluded that the L2 students in Hong Kong tended to avoid overt author identity when it involved making arguments or claims. The relationship between writing competency and authorial voice,

apparently, is still very open to debate in the writing research. On the basis of an empirical study, Helms-Park & Stapleton (2003) first suggested that there might not be a relationship between construction of individualised voice (use of first person singular as one variable) and the quality of writing (determined by rating). However, Zhao & Llosa's study (2008) contradicted Helms-Park & Stapleton's (2003) by means of the same methodological design. In comparison with the finding in the current study, it might be argued that the use of authorial voice is very likely to hinge on the text genre. The impersonal pronoun *it* is preferred in academic writing, particularly in science subjects, in order to present an objective and impartial perspective. In the context of general argumentative and expository writing that learners most often engage in, writer/reader visibility might not be directly related to whether it is a piece of good writing. Rather, learners should be advised to use personal pronouns in a moderate and appropriate manner.

7.3 'Foreign-soundingness' in the Use of Lexical Bundles

The first section in this chapter set out from the structural and functional distribution of lexical bundles with the aim of describing the amount of difference and/or similarity of bundle use among the five groups of writing. A more in-depth investigation suggested that the quantitative analysis needed to be complemented by closer examination of lexical bundles under the same category in each group and also by scrutiny of concordance lines. By taking advantage of such a hybrid methodology, a number of distinctive discourse features varying with levels of writing competency have been identified. We have seen that the learner writers show infelicities in various stylistic and pragmatic aspects such as a lack of sensitivity to the spoken/written register difference when compared with native expert writing. The infelicities demonstrated in the use of recurrent strings in learner writing may be partly attributable to the current ESL/EFL teaching syllabus and also partly to L1 specific transfer as well as universality in L2 acquisition. In this section, therefore, I wish to explore different possible

explanations behind the 'foreign-soundingness' or 'non-nativeness' in learner writing.

One kind of word combinations found in the learner data, bundles which contain the existential structure '*there is/are* + NP', is probably specific to L1 Chinese learners, and this learner idiosyncrasy is particularly pronounced at the lowest level (CEFR-B2 writing). As discussed in Section 6.5.4.1., in many instances of the '*there is/are* + NP' bundles found in CEFR-B2 writing, this existential structure is often erroneously followed by an infinitive, a finite verb or a relative clause. This might be due to a similar construction in Chinese, '有 (yǒu, *there is/are*) +NP', which allows the existential 有 (yǒu) to precede a verb phrase. Aside from this specific case, many stylistic or pragmatic infelicities found in learner writing appear to be more universal for second language learners. Take the colloquial elements in learner writing for example. Gilquin & Paquot (2007) investigated data from 14 L1 populations in the International Corpus of Learner English (ICLE). They outlined various spoken features commonly shared by learners of different L1 backgrounds (e.g. the use of adverb *besides*, the exemplification *like*, and sentence-initial *and*). Similarly, as discussed in Section 7.2.4, the tendency to make overgeneralisations is not exclusive to L1 Chinese learners (cf. Altenberg & Tapper, 1998; Kennedy & Thorp, 2007; Ringbom, 1998).

In addition to L1 transfer and L2 acquisition universality, ESL/EFL teaching syllabi might also have a greater impact on the idiosyncrasies in learner writing. It has been reported that learners tend to overuse certain types of lexical items while underusing others (Altenberg & Tapper, 1998; De Cock, 2000; De Cock, Granger, Leech, & McEnery, 1998; Granger, 1998; Granger & Rayson, 1998). A preliminary examination of a few available EAP course books or manuals for academic writing indicates that the overused expressions in learner writing, generally the structurally and semantically complete ones, are usually among lists of

connectors or tables of linking phrases in the published materials.[47] The expressions *on the other hand* and *at the same time* are two typical examples. The former is taught in a few of the materials examined (Jordan, 1999; Oshima & Hogue, 1999; Swales & Feak, 1994) while the latter is found in only one course book (Jordan, 1999). Despite their distinct degree of spread in the materials, it can be seen from Figure 7-13 that the learners at various proficiency levels used these two formulae much more frequently than native speakers.

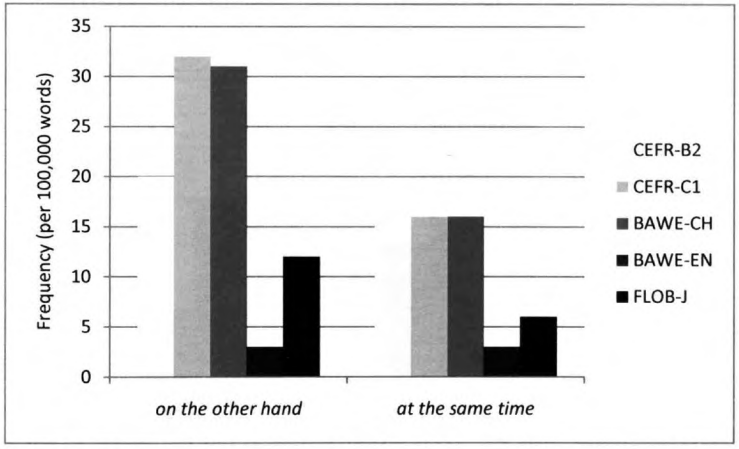


Figure 7-13 Relative frequency of *on the other hand* and *at the same time*

---

[47] It is not intended in this thesis to make a thorough survey of all the learning materials for academic writing. The course books or style manuals examined, therefore, are those available from the Lancaster University Library. Only those which were found to be related to the discussion here are listed below:

Rosen, L. J. (2008). *The Academic Writer's Handbook* (2nd ed.). Harlow: Longman.

Bailey, S. (2006). *Academic Writing: a Handbook for International Students* (2nd ed.). London: Routledge.

Jordan, R. R. (1999). *Academic Writing Course: Study Skills in English* (3rd ed.). Harlow: Longman.

Oshima, A., & Hogue, A. (1999). *Writing Academic English* (3rd ed.). White Plains, NY: Longman.

Swales, J., & Feak, C. B. (1994). *Academic Writing for Graduate Student: Essential Tasks and Skills: a Course for Nonnative Speakers of English*. Ann Arbor: University of Michigan Press.

We now turn to other lexical bundles which are among the top most frequent expressions in native writing but rarely or never mentioned in the learning materials. As exemplified in Figure 7-14, *in terms of the* and *in the case of* are two of the most frequently used bundles in FLOB-J; they were, however, used far less frequently by the learners in BAWE-CH. In effect, they were not even in the bundle repertoires of CEFR-B2 or CEFR-C1 writing.[48]

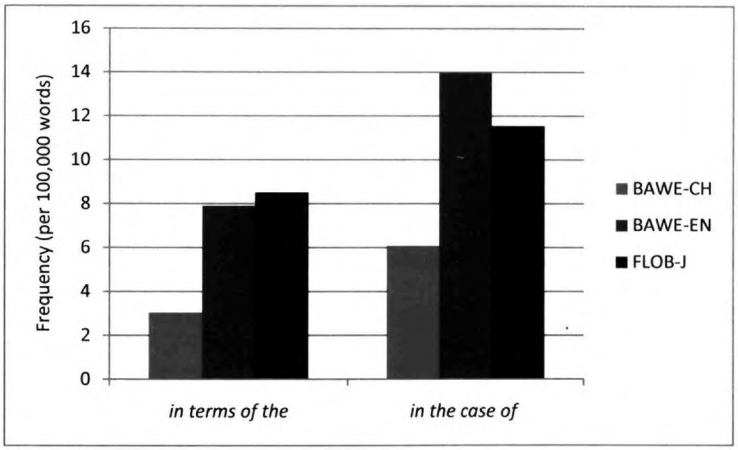


Figure 7-14 Relative frequency of *in terms of the* and *in the case of*

It might be proposed that these highly frequent bundles in native writing such as *in terms of the* are not so much developmental bundles as academic clusters as the genres in Modular Studies 1 and 2 are not exactly the same. As addressed in Chapters 4 and 5, the two corpora investigated in Modular Study 1, FLOB-J and BAWE, are defined as being academic writing for they were produced in an academic context and their content was built upon areas of knowledge such as physics or economics. By comparison, the majority of the texts

[48] There are instances of *in terms of the* and *in the case of* in CEFR-B2 and CEFR-C1 subcorpora, but they failed to reach either the frequency or the dispersion thresholds (at least four times in three texts or more).

investigated in Modular Study 2, i.e. argumentative and expository essays from the Longman Learners' Corpus, would be categorised as EAP-like essays (English for Academic Purposes), which mainly focus on the ability of English language as opposed to any subject content. The rationale is that argumentative and expository essays are generally the text types targeted in EAP courses as academic writing is often mixed with argumentation as well as exposition in nature. Such a genre variation might have had an effect on the use of lexical bundles. Yet it is evident that learners are not exposed to these academically essential clusters in the learning materials. It is very possible that the lack of explicit learning has resulted in the low frequency of 'academic' or 'literary' clusters or even the absence of them in learner writing. In addition, lexical bundles are a mixture of various forms of phraseological units. We have seen that many of the bundles are fragments while others are not. The attributes of certain types of bundles might also influence the acquisition and usage of them in learner writing.

It has been generally acknowledged that there exists a continuum of the phenomenon of phraseology ranging from the most fossilised word combinations to the least fixed ones (e.g. Cowie, 1998; Howarth, 1998b). For example, Howarth (1998a) discussed different ways that verbs can be collocated with objects. This can range from purely idiomatic cases (e.g. *blow the gaff*) to free combinations (e.g. *blow a trumpet*). A similar relationship has also been observed in the use of lexical bundles in learner writing between frequency and the gradient span of fixedness and explicitness of bundles. At one end of the dichotomy are the expressions with holistically semantic and structural salience (e.g. *on the other hand*, *at the same time*, *all over the world*). These bundles appear to be quite fossilised, and no components within the bundles in question can be replaced by other equivalent units, which means that they do not constitute a productive 'fixed frame' with other bundles with a similar structure (termed as 'phrase-frame' by Stubbs, 2007) (cf. Section 3.2 and Section 5.3.4.1). These formulae are largely seen in the ELT materials and are often overused by learner

writers (cf. the clichéd bundles discussed in Section 7.2.2). On the other end are expressions which appear to be less holistically salient in terms of semantics and structure (e.g. *in the context of*, *the way in which*, *it is clear that*). These bundles are usually used to bridge two structures, and very often they can be categorised under a 'fixed frame' such as '*the* + Noun + *of the/a*' or '*it is* + Adj.+ *that*', which are rarely or less often discussed in the ELT materials although many of them are highly frequent in native expert writing. Interestingly enough, the holistically explicit expressions are often among those most frequently used word combinations in learner writing. It is possible that explicitness results in these expressions being included in ELT instruction and hence L2 learners overemphasise them to the degree that their writing begins to sound unnatural or 'foreign'. In contrast, those less explicit chunks, usually perceptually incomplete ones, are highly frequent in native expert writing but rarely used by learner writers as they might have been overlooked in ELT instruction and learning materials in the past. Surely the above discussion can only provide the explanation to some overused and underused bundles because not all the overused bundles are perceptually salient (e.g. *is one of the*) and not all the underused bundles are perceptually fragmental (e.g. *on the one hand*).[49] It is also very likely that such a dichotomy is a scalar array rather than a binary phenomenon, which is worth more research in the future. In short, the above finding has some implications for second language pedagogy regarding how to balance the overuse and underuse of the formulaic expressions.

On the basis of the above discussion, at least three issues can be addressed for

---

[49] In the keyness analysis discussed in Section 7.1.3, *is one of the* is defined as one of the overused bundles in all the four student or learner groups when compared with expert writing in FLOB-J. In contrast, although *on the one hand* is not highlighted in the keyness analysis, it occurs eight times in FLOB-J but is not listed in any of bundle repertoires in the non-expert groups of writing. It appears that native academics tend to make use of the pair of *on the one hand* and *on the other hand* whereas students or learners tend to use *on the other hand* alone.

ESL/EFL teaching and learning. Firstly, it is necessary to broaden the range of formal/academic expressions available to the learners by adding more clusters which have been overlooked in the past, particularly the less explicit expressions. Secondly, the importance of moderate use of certain expressions, particularly the overused ones (e.g. *on the other hand*), has to be emphasised. Lastly, learners should be advised that the discourse features found in learner writing, including clichés, verbosity, colloquialism, and overgeneralisation, are undesirable in formal/academic writing and should be avoided.

Recently, some learner-corpus researchers have been engaged in putting the research findings into practice in the area of language learning. Gilquin, Granger, & Paquot (2007), for instance, described a collaborative project with Macmillan Education, in which the research team incorporated their research results from the ICLE corpus they have been working on for the past decade with the publication of EAP materials. This project is exemplary in the sense that very few EAP materials have incorporated learner corpus data to inform the contents. In the end product of this project, the *Macmillan Dictionary for Advanced Learners* (Rundell, 2007), a number of rhetoric functions that learners often have difficulty with have been highlighted with 'Be careful!' notes, 'Get it right!' boxes, and frequency graphs aiming to help learners reach native-likeness. However, despite the attempt to be comprehensive, the dictionary still fails to point out a few features distinctive in native academic writing. For example, the two productive frames, '*the* + Noun + *of the*' and '*in the* +Noun + *of*', are characteristic of native expert writing, but they are not introduced. Additionally, Gilquin et al. (*ibid*) still largely focused on word-level and phrase items (e.g. *really*, *of course*, *on the contrary*) rather than the less perceptually salient structures which are highly frequent in academic discourse (e.g. *the extent to which*, *the way in which*, *in the case of*).

Some of the stylistic features discussed in this thesis actually have been included in several EAP materials. For example, in addition to connectors, the use of cautious language

has also been extensively touched upon (e.g. Bailey, 2006; Jordan, 1999; Swales & Feak, 1994). What is lacking in these EAP materials is the supplementation of authentic examples contrasting the usage between the formulaic expressions preferred in the native norm and the overstated statements generally observed in learner writing from corpus data. Meanwhile, verbosity or cliché, also prevalent in learning writing, is rarely referred to in EAP materials. Including warning boxes or reminding remarks along with examples would help learners distinguish and corroborate these core stylistic notions in their writing.

It has to be stressed that the investigation of EAP and academic materials in this thesis is not comprehensive and therefore cannot be claimed to be representative. It is beyond the scope of this thesis to compare individual lexical bundles against the full range of ELT materials available. Yet it would be interesting to carry out a systematic survey to investigate the extent to which the gap exists between learner writing and EAP syllabi in terms of use of phraseology and rhetoric guidelines. In fact, a relevant survey of this kind has just been conducted by Bennett (2009), with the focus on general features in academic writing such as text structures, grammatical issues, and lexical features. In spite of the variations in readership, genre and discipline demonstrated in the 40 style manuals she investigated, Bennett found that the academic discourse in English established in these manuals displayed a remarkable degree of consistency. This finding is very encouraging for EAP writing pedagogy in the sense that learners would only need to follow one set of general principles and features desirable in the academic discourse rather than adjusting to various disciplines or genres.

Another caveat to be born in mind is that some of the distinctive formulaic language revealed through comparing the five groups of writing may be the result of *proficiency development* as well as *task difference*, which can be difficult to distinguish apart. The use of cautious language, for example, would not be as important in an argumentative or expository

essay as in a journal scientific paper. There appears to be a subtle and complicated interaction between writing competency and genres. While the five groups of writing are considered to represent writers of different proficiency, the text types in each group also respond to diverse task requirements. It has to be acknowledged that the comparison made in this thesis is a compromise: EAP-like argumentative or expository essays rated and chosen from the Longman Learners' Corpus and university assignments versus published academic prose. In the real world where linguists collect naturalistic data to investigate L2 writing development, certain genre types can probably only be produced by writers who have reached a certain proficiency level. That is to say, certain genres require a higher proficiency level than others. It would be virtually impossible for an L2 writer at the entry level, say a CEFR-A2 writer, to produce an argumentative essay. Likewise, it would be equally impossible for a CEFR-B2 writer to write a scientific paper aimed at publication in a journal. Meanwhile, in the course of learning to write for academic discourses, when novice writers progress and arrive at a certain level, they usually focus on producing the kind of writing more or less corresponding to their proficiency level, at least according to the general ELT syllabuses. Such a relationship between text genres and learner proficiency can be demonstrated from the writing tasks used in the Cambridge Main Suite exams and IELTS. In the exams for lower levels (e.g. PET), we see writing tasks such as writing a short message or a 50-word story. Moving on to the intermediate level and above (e.g. FCE and CAE), learners are required to produce a report, a proposal, an essay or a review. Targeting L2 learners who plan to study in an English-speaking country, the IELTS Academic[50] test requires learners to complete two types of writing tasks: one is analytical writing making use of the information given by visual input such as a graph (at least 150 words), and the other an argumentative essay (at least 250

[50] There are two modules of IELTS tests: general training and academic. The writing tasks are different in these two versions of the test.

words). The various task types reflect that learners are generally expected to begin with narrative or descriptive which contain more colloquial language and then continue towards argumentative or expository literacy with increasing text length. Learner writing appears to be a process starting with orality and ending in literacy. In second language research, therefore, it is virtually unachievable and probably undesirable, to acquire the naturalistic data which are contributed by learners at different levels responding to identical writing tasks. Accordingly, while the use of formulaic language is compared across writing of various proficiency levels yet different genres, the conclusions have to be tackled with caution.

The above notion might help to explain to some extent why the lower level CEFR-B2 writing would appear to contain more speech-like elements than other groups as the CEFR-B2 level is probably at the transition stage where learners begin to grasp the distinction between formal and informal writing. Interestingly, if we refer back to the CEFR Written Assessment Criteria used as the rating scale in the current research (See Appendix 3), the CEFR-B2 writers are described as being able to 'make a distinction between formal and informal language with *occasional* less appropriate expressions', and their 'language lacks, however, expressiveness and *idiomaticity* and use of more complex forms is still *stereotypic*' [emphasis added]. It is felt that on the basis of the findings in this thesis, the extent of informality discovered in CEFR-B2 writing appears to be more severe than simply occasional inappropriacy (e.g. the undue use of bundles with the colloquial quantifier *a lot of*). This tendency to be speech-like, nevertheless, is not found in the lexical bundles in CEFR-C1 and BAWE-CH writing. By contrast, the lack of native idiomaticity and the stereotypicality in the use of certain lexical bundles is not only marked in CEFR-B2 writing but also lingering in CEFR-C1 and even BAWE-CH writing (e.g. the preference of certain formulae *on the other hand, all over the world*, and the absence of two frames '*the* + Noun + *of the*' and '*in the* + Noun + *of*'). Yet such features are not seen in the CEFR assessment criteria grid at the levels

above CEFR-B2, and the descriptors regarding styles or formulaicity are rare except for the one for CEFR-B2 quoted above. The only descriptor which can be remotely related to the discussion here is the statement found in CEFR-C1: 'The flexibility in style and tone is somewhat limited'.

It also has to be stressed that using rated essays to investigate second language development is by no means a circular practice, as some might claim. Performance rating is a complex judgment process, in which various characteristics could all impact on the measurement (e.g. for difference between native and non-native raters, see Kim, 2009; for the use of rating scale, see Kondo-Brown, 2002). In the case of adopting a CEFR rating scale here, as already stated, the notion of formulaicity or idiomaticity is rarely addressed in the whole assessment criteria grid except for the brief statement just mentioned above regarding the CEFR-B2 level. The results in this thesis can thus provide some empirical underpinnings for not simply a rating scale but also a large-scale framework of reference for languages such as CEFR.

7.4 Summary

In this chapter, we have seen an overview of both the quantitative and qualitative analyses of bundle structures and functions in Modular Studies 1 and 2. Despite the genre difference of these subcorpora investigated, learners appear to share some similarities in terms of the use of lexical bundles across proficiency levels and task types within learner writing only (e.g. the tendency to make overgeneralisations and use more clichés). Native speaker groups, on the other hand, also used some bundles which are not present in learner data (e.g. most hedging devices or extent/degree modifiers). Coupled with chi-square standardised residuals, the structural and functional categories which can better distinguish different groups of writing proficiencies have been compared and identified, although the results are not always consistent throughout two studies. Then via the keyness analysis, some aspects which had

been overlooked in the quantitative analysis and qualitative examination have emerged (e.g. *is one of the* being overused in all the four non-expert groups of writing). Such a combination of multiple approaches in investigating lexical bundles has proved to be very effective in searching for distinctive features across writing competency. It has been found that at the lower proficiency levels, learner language tends to be more simplistic, colloquial, clichéd, verbose, overstating, and the authorial voice appears to be more personally involved while the more proficient writing tends to be more native-like, thereby demonstrating an opposite pattern. The implications for second language pedagogy, assessment, and psycholinguistics will be further discussed in the next chapter.

Chapter 8 Conclusion

In this chapter, the conclusion drawn from analysis will start with the theoretical status of lexical bundles in comparison with that of conventionally defined formulaic language. The limitations of current research will then be foregrounded and discussed, followed by the implications for various related areas such as second language research and language testing. Directions for future research will also be addressed. Then this chapter will end with some concluding remarks.

8.1 Status of Frequency-Driven Formulaic Language in SLA

As explained at the beginning of this thesis, two kinds of approaches can be distinguished in linguistic studies which make use of a large collection of text data: corpus-based and corpus-driven (Tognini-Bonelli, 2001). Similarly, a pair of contrasting notions have been employed in defining phraseology: the conventional linguistic analysis of various kinds of multi-word units and the more recent statistical or frequency-based approach to identify lexical co-occurrences (Granger, 2005). The conventionally defined formulaic language is usually researched with a pre-existing repertoire of formulaic expressions such as Moon's study (1998b) or using native-speakers' judgments to identify the formulaic sequences such as Li & Schmitt's longitudinal study (2009). This conventional approach appears to encompass various kinds of formulae, including proverbs (e.g. *better late than never*), phrasal verbs (e.g. *get up*), fillers (e.g. *you know*), similes (e.g. *as white as a sheet*), metaphors (e.g. *ring a bell*) among others. This top-down approach of signposting word combinations is often characterised by non-compositionality as one of the criteria to categorise the phraseological units (Cowie, 1998; Howarth, 1998a, 1998b). Yet there are two potential problems with those traditionally perceived fixed expressions. On the one hand, the notion of compositionality/non-compositionality per se is complex and controversial enough as an 'all-

embracing' criterion for fixed expressions and prefabs because certain defining features of compositionality involves scalar judgment rather than either/or (Svensson, 2008). As a result, it still remains unclear to what extent the researchers can have confidence in demarcating the boundary for phraseology and distinguish different types of phraseological units. On the other hand, these traditionally defined formulae do not occur as frequently as we might have expected. Moon (*ibid*) compared a list of 6,700 fixed expressions extracted from *Collins COBUILD English Language Dictionary* against the 18-million-word Hector Corpus, the majority of which is journalism and other non-fiction texts. Surprisingly, she found that 70 per cent of the expressions fail to occur more than once per million words in the Hector Corpus. That is to say, over two-thirds of the phrase entries in the dictionary actually occur with minimum frequency in real-life data. Moon estimated that the frequencies of five or more per million words would enable us to be 'slightly more confident of observing them again in other broadly comparable corpora' (cf. the normalised cut-off frequency of determining four-word bundles in this thesis: 25 and 45 times per million words). The finding in Moon's study poses a serious question in SLA which is worth our consideration: if these idioms, formulae, or whatever multi-word units do not occur that often, are they worth as much attention as in the current ESL/EFL teaching syllabi?

The concept of frequency-based formulaic language might be able to provide a remedy for the above problems and open up a new prospect for second language teaching and learning. With the development of corpus analytic tools, some linguists have begun to view word combinations from a bottom-up perspective. The multi-word units retrieved by this corpus-driven approach, either by frequency or more complex statistical calculations, are usually compositional, not necessarily holistic units. As the current research reveals, such an automated inductive approach has shown that the majority of recurrent word combinations in native academic prose are NP-based and PP-based bundles in terms of structures and

referential expressions in terms of functions. In addition, certain frames such as '*in the* + Noun + *of*' or '*the* + Noun + *of the*' are highly productive, which have not yet received much attention in ESL/EFL and EAP writing instruction. Different from the conventional approach, the recurrent phraseology often functions as 'building blocks' in the texts, organising the textual discourse. The rationale behind this approach is that if these recurrent word combinations are used repeatedly with the same fixed order and the same form by different speakers/writers, it is very likely that they are stored and processed as a single unit, instead of novel combinations generated under some sort of rule-based grammar right at that moment. As Biber & Barbieri (2007, p. 284) point out, *frequency* and *salience* have been argued as two parameters affecting the acquisition of language features. L2 learners might benefit to a large extent from the perceptual salience of conventionally defined formulaic language. Yet the highly frequent lexical bundles, despite being largely less perceptually salient, can also help learner writers to achieve a more native-like formulaic style. We have seen the power of frequency in the earlier analysis. Although the frequency-defined formulaicity embodied through a selection of recurrent word combinations constitutes merely 1-2% of the running words, it has cast some light on the distinctive discourse features which can be used to distinguish writing competency and also have great implications for language pedagogy.

What second language education needs, as Granger (2005) indicates, is the reconciliation of these two contrasting approaches, i.e. integrating the traditionally recognised phraseological units and the frequency-driven ones into SLA teaching practice. By comparing the frequency-driven phraseological units in groups of writing, this research has pinpointed a number of learner idiosyncrasies in the use of lexical bundles across various stages of development. It is hoped that in future research, the findings here can be compared with the results from using the conventional approach and thus can make some contribution to this reconciliation.

8.2 Limitations

Certainly, there are a few inevitable limitations in this thesis, and they are mostly concerned with the accessibility to corpus data and the comparability of data. First and foremost, the written corpora representing different proficiency levels compared should ideally be of similar size and composed of preferably identical or at least parallel writing tasks with similar length. In reality, as discussed in various sections earlier (see Sections 6.2 and 7.3), perfect data comparability is virtually unachievable, particularly in non-experimental contexts. The imbalance of constituents between each subcorpus, e.g. the number of texts and average text length, is certainly undesirable. Yet this does not necessarily invalidate the findings in this thesis. Reassuringly, Oakey (2009) compared 'isolexical' and 'isotextual' versions of a corpus composed of eight subcorpora from different disciplines, the former with an equal number of tokens (but not texts) in each subcorpus and the latter with an equal number of texts (but not tokens). He found that text length does not actually affect the retrieval of lexical bundles. Additionally, the examination of lexical bundles indicates that only a couple of word combinations appear to be clearly related to number of texts (e.g. *in the following paragraphs* in CEFR-B2 and *in this essay I* in BAWE-EN).

What is more worrisome than text length and number of texts, however, is the genre variability across the subcorpora that were compared. With the advance of proficiency, learners confront various requirements of writing tasks corresponding to their proficiency. It is unlikely that researchers will be able to collect perfectly comparable data with exactly the same genres from writers of different proficiencies. Certain linguistic features might only be triggered in certain text types (e.g. *as a function of* and *the per cent of* in scientific writing found in FLOB-J). The best compromise is probably to compare the texts which are considered to be at least in the same dimension, i.e. comparing published academic writing and university assignments in one study, and EAP-like argumentative and expository learner

writing across proficiency levels in another, although there are still more constraints in each modular study.

For Modular Study 1, it has to be acknowledged that the use of FLOB-J as the representation for native expert academic writing might have had some impact on the word combinations derived. First of all, a large body of the texts included in FLOB-J are hard-science based. This is probably why we found bundles such as *a function of the*, *the magnitude of the*, *the structure of the*, and *a high level of* in FLOB-J, which appear to be strongly concerned with the disciplines of hard science. Meanwhile, the journal papers or book sections selected in FLOB-J are all 2000-word long excerpts rather than complete texts (which were what the BAWE student writing corpus was comprised of). It is likely that there might be more occasions in the BAWE student writing to use discourse organisers as student essays are mostly structured as Introduction, Body, and Conclusion. However, when examining the concordance lines, very few discourse organisers were found which could possibly attribute to the difference of excerpts and full texts, i.e. topic-introduction bundles such as *in this essay I* from BAWE-EN or *last but not least* from BAWE-CH.

For Modular Study 2, one potential limitation is concerned with the impact from task variation upon rating. Although the whole process of rating adopted is considered to be fairly robust following a set of standardised procedures recommended by the official CEFR manual, the variability of writing tasks might have affected the raters' judgment. In normal practice of language testing, raters are often required to mark learner performance elicited by identical or parallel tasks. By contrast, the data extracted from the Longman Learners' Corpus includes various types of writing tasks ranging from university term papers to general learner essays (cf. Section 6.3.4). Great efforts, however, have been made to cover as a wide range of task types as possible in benchmarking and rater training with the aim of minimising the impact from task variability.

The last constraint related to the data is the fact that the learner writing from both the Longman Learners' Corpus and the L2 component from the BAWE corpus is not error-tagged. We cannot know for sure if any learner errors might have affected the generation of word combinations. As explained in Section 7.1.5., it is very possible that the deviant forms found in learner data, owing to the nature of unpredictability, would result in under-representation of certain word combinations. From another perspective, however, error tagging has always been a notoriously difficult task. The literature review in Section 2.1.1 has shown that only a few error-tagged learner corpora are available (e.g. the ICLE or Cambridge Learner Corpus) and that error-tagging very often requires a complicated annotation scheme which would take a great amount of time and effort. Yet as demonstrated in this thesis, taking advantage of the automated approach of retrieving word combinations without any error annotation can still generate some insightful results with regard to distinctive discourse features in learner writing.

Apart from the data, one constraint that might affect the validity of findings is the consequences of this automated approach. This frequency-driven methodology apparently cannot cater for discontinuous word combinations, and thus certain information might be missing. Additionally, only identical forms of word combinations can be computed; therefore, the inflectional change of a noun or verb form can not possibly be automatically included and combined (e.g. *the way in which/the ways in which* in the case of noun inflection, and *have a lot of/has a lot of* in the case of verb inflection). Other variations of a base form, say the insertion of an adverb in a lexical bundle (e.g. *it is difficult to* and *it is very difficult to*, cf. Section 7.1.5), also impact on the degree of preciseness in quantitative analysis. It is possible to inspect those variations and combine them manually, yet the problem is to what extent and how thorough this may be achieved as there can be all sorts of variations for the same set of word combinations (e.g. insertion of an adverb, a negation word, etc.). Researchers who work with this approach, therefore, have to be cautious of the consequences of under- or over-

representation of this sort, particularly when conducting quantitative analysis.

8.3 Implications

This section will start with discussion of methodological issues involving determining and comparing lexical bundles across corpora as well as considering issues relating to rating learner essays. Following that, I will discuss the implications of the study for psycholinguistics, second language teaching and learning, SLA research, and language testing.

8.3.1 Methodological Issues

Three subcorpora were compared in Modular Study 1, and another two subcorpora in Modular Study 2. The average size of each corpus in Study 1, however, is nearly twice as large as that in Study 2 (approximately 155,000 words versus 88,000 words), although the same threshold for defining lexical bundles was adopted, i.e. word strings occurring at least four times in three texts or more (thereby around 25 and 45 times per million words in Modular Studies 1 and 2 respectively). As illustrated in Section 3.5, the decision to use an identical threshold with raw frequency was made on the basis of repeated experiments with the corpus data, and the quantity and quality of retrieved word combinations were found to be appropriate for the scope of this thesis. Different from the literature, it is also argued in this thesis that both the raw cut-off frequency and the corresponding standardised frequency should be reported in order to transparently reflect the methodology adopted. The claim that using an identical normed threshold, such as 20 or 40 times per million words for each of the (sub)corpora investigated, might sound 'impartial' in the beginning. Yet after the standardised rate is converted to raw frequencies, it could substantially affect the number of generated word combinations when comparing corpora with various sizes. As we have seen in Section 3.5, with the cut-off standardised frequency set at 40 times per million words, the converted raw frequency threshold for a 5.3-million-word corpus is as high as 212 times whereas for a 40,000-word corpus, the converted rate is far much lower at 1.6 times. Adopting a

standardised frequency threshold to compare corpora with various sizes, as contended by Biber and Barbieri (2007), does not influence the conclusion drawn on analysing the bundles when complemented with adjustable distribution requirements. It could be, however, misleading to report that a standardised frequency criterion for corpora of various sizes is the only fair solution when extracting recurrent word combinations, because the converted raw frequency does not reflect such fairness.

The second methodological issue concerns the numbers of lexical bundles compared across corpora. In the research examined in the literature review (Section 3.3), we see 19 bundles retrieved from academic prose compared with 84 bundles representing classroom teaching in Biber, Conrad, and Cortes (2004) or 19 academic-prose bundles compared with around 130 bundles from written course management. When Biber and his colleagues investigated how the discourse functions of these lexical bundles qualitatively differ across registers rather than carrying out a quantitative analysis, the drastic difference of quantities of bundles compared did not necessarily undermine their analysis. Yet I would like to argue that in order to present a more comprehensive scope, more lexical bundles should be sought for investigation as opposed to only 19. Considering the comparability with other studies, 100 bundles, more or less, seem to be a good number for each corpus investigated. Surely this requires repeated experiments with the corpus data so as to figure out optimal frequency and dispersion thresholds. As discussed, the quantity of meaningful recurrent word combinations (i.e. those after filtering and refinement) representing frequency-driven formulaicity (types) in our mental lexicon appears to be finite while their occurrences (tokens) are infinite increasing with corpus size. In theory, as the corpus grows larger, we are more likely to retrieve a wider range of different bundles. The increase rate of the number of bundles with growing corpus size, however, is not linear. I suspect that after a certain critical point, the increase rate of number of bundles would begin to level off. Only when we figure out the

relationship between corpus size and the optimal determining thresholds will we be able to make any assertion with regard to how the number of lexical bundles interacts with proficiency as Hyland (2008a) or De Cock (2000) proposed.[51] Seeking out this critical point with the most appropriate cut-off frequency and dispersion threshold could be one direction for future research.

During the process of experimenting with various cut-off frequencies (cf. Section 3.5.1), it was also found that raising the threshold, i.e. adopting a much stricter criterion, would generate a set of 'neater and cleaner' bundles with less overlapping and less context dependence, but the researcher also ends up with less information. With a lower threshold, more word combinations would be retrieved which would afford researchers more data for investigation, but the undesired overlapping and context dependence also occurs more often, which means that the researcher needs to spend more time and effort to tidy up the data. Determining a threshold is thus a tug of war between the amount of information required and the degree of precision/representativeness favoured.

Another methodological issue arises from categorisation of lexical bundles. In terms of structural categorisation, we have seen that a few lexical bundles are assigned to the structural categories which do not reflect their syntactic role in a sentence/clause (see Section 4.2.3). To be more specific, *that is to say* is categorised under VP-based bundles while *on the other hand*, *at the same time*, and *all over the world* are categorised under PP-based bundles, although they mostly function as adverbials in the contexts where they occurred. This is one potential problem with the structural taxonomy that has to be acknowledged, although such adverbial bundles are few and far between. In addition, the existence of the 'Others' category

---

[51] Other potential problems in Hyland's (1008a) and De Cock's (2000) studies is that they did not deal with context dependent bundles and overlaps prevalent in the retrieved word combinations, which would presumably invalidate the quantitative comparisons they made.

(including six bundles: *as long as the*, *as soon as the*, *as well as the*, *last but not least*, *than that of the*, and *whether or not to*) also suggests that the categorisation scheme is not perfect and that other structural categories of bundles could be created if more instances with a similar structure can be found. In terms of functional categorisation, some might be more sceptical about to what extent the functional categorisation is valid and reliable when quite a few lexical bundles actually hold more than one discourse function. For example, *this may be due to* is assigned to the category of inferential text organisers, but it can also be a stance epistemic expression since it contains a modal element *may*. Some judgments on categorisation unavoidably can be regarded as being more or less subjective, and it is possible that one might disagree with some of the assigned categories, which would be a common problem to almost all forms of categorisation. The key, however, is to be consistent and transparent. In effect, as discussed in Section 4.3.1, Biber and his research colleagues (2003, 2004, 2007) kept changing the allocation of a couple of functional subcategories. In an ideal research setting, two annotators should be involved with this categorisation task, and the usage of each concordance instance of lexical bundles should be documented before a final classification decision can be made. Then with a high reported inter-rater reliability, we can claim with more confidence that the determination of bundle categorisation is justified and well grounded. In reality, very few studies can afford such a luxury of required time and manpower. Hyland's solution for the ambiguity concerning overlaps of categorisation was to examine a large sample of instances which constitute around 17% of the total occurrences (2008a, p. 49) to ensure that there was correspondence between the functional category assigned and the context of occurrences. From the experience drawn upon classification bundles in this thesis, I agree that assigning a discourse function for one bundle should accord with the majority of its occurrences, but this does not solve the problem that one expression may serve more than one function as the example of *this may be due to* mentioned above. I

would like to posit that as long as a categorisation system is established on a sound framework with explicit definitions and illustrative examples, then consistency and transparency is the only key to effective categorisation. After all, a quantitative analysis still requires the complement of qualitative examination as has been demonstrated in this thesis so as to provide a better overview of how the use of lexical bundles differs across development. In the end, the categorisation is but a means to enable researchers to carry out a more systematic analysis so that some aspects would not be overlooked if only a comprehensive list of clusters is considered.

To sum up, as determining lexical bundles is a frequency-driven approach which directly affects the quantity and quality of the word combinations retrieved and investigated, the definitive variables such as corpus size, cut-off frequency and dispersion, numbers of word combinations, ways of categorisation, should all be taken into account before any valid claims from the analysis can be stated. It is hoped that this thesis has thrown some light on our understanding of these methodological issues which can benefit future research.

8.3.2 Psycholinguistics

A large number of studies have been dedicated to research of formulaic language in the spoken register (cf. Section 3.2), most of which are concerned with the socio-interactional functions that are generally associated with spontaneous speech. It has also been found that speech contains more formulaic expressions than writing (e.g. Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2003). However, as evidenced by this thesis, academic writing also embraces a diverse range of formulaic expressions. The function of formulaicity in a language can be interpreted via two kinds of perspectives (see a thorough discussion in Wray, 1999; Wray & Perkins, 2000). On the one hand, formulae are the preferred ways of expressing something when there are a great many other possible word combinations available. By choosing the most recognised and accepted form among all of the

grammatically possible options, the speakers/writers can be identified as belonging to a specific discourse community. Second language learners who want to achieve proficient academic writing have to learn how to use academic formulaic clusters in the same way as native academics. On the other hand, formulaic language is often interpreted by psycholinguists as multi-word units which are stored and processed holistically as opposed to compositionally. The processing benefits are two-way: for both the producer (speaker/writer) and the receiver (hearer/reader). The above two perspectives are actually two sides of one coin. The preferred ways of saying or writing something suggest that formulaic language is more likely to consist of prefabricated units retrieved from our mental lexicon rather than novel combinations generated from a rule-based grammar when confronted with a lexical selection. However, is there any psycholinguistic evidence to support this holistic view of corpus-driven formulaic language?

It has to be stressed that the hypothesis of holisticality and processing reduction for formulaic language is still disputable when the formulae are extracted from frequency-driven and corpus-based studies and then tested with a variety of psycholinguistic experiment settings. As addressed in Section 3.1, Schmitt et al. (2004) found that not all corpus-derived clusters were psycholinguistically valid whereas Jiang & Nekrasova (2007) concluded that their findings provided prevailing evidence in support of the holistic nature of formula representation and processing in both native and non-native speakers. This inconclusiveness, nonetheless, probably truthfully reflects the nature of frequency-derived word combinations, i.e. a mixture of various kinds of highly frequent formulaic sequences ranging from the most fixed formulaic sequences (e.g. *last but not the least*, *on the other hand*) to the seemingly least idiomatic ones (e.g. *it is possible to*, *can be used to*). Those at the fixed end of corpus-derived clusters are generally self-contained and perceptually salient while those at the least idiomatic end are often the fragments bridging two units, which are structurally and

semantically salient to a much lesser degree. As discussed in Section 7.3, a dichotomy of lexical bundles which concerns perceptual salience and holisticality was proposed for attribution of L2 learners' overuse and underuse to over-emphasis of certain types of bundles or complete lack of other types of bundles in ESL/EFL instruction. A similar dichotomy can be applied for the holistic theory of formulaic language. Again, it is quite unlikely that the psycholinguistic mechanism of formulaic sequences is a binary phenomenon, i.e. stored and processed either *holistically* or *analytically*. It is more probable that the holistic mechanism functions with a gradient, depending on the extent of structural and semantic salience of word combinations. It is also possible that the mechanism hinges on how frequently the clusters recur in our language. In theory, the more frequent they are, the more likely they would be processed holistically so as to reduce the processing load.

Interestingly, quite a few psycholinguistic studies working on the theory of processing advantage over formulaic sequences other than corpus-driven clusters also show conflicting findings, with some supporting the holistic view and others not (e.g. Conklin & Schmitt, 2008; Schmitt & Underwood, 2004; Underwood, Schmitt, & Galpin, 2004). The issue is further compounded by the use of different operationalisations for defining formulaic language in various studies (e.g. so-called 'idioms' extracted from dictionaries) combined with various psycholinguistic experimental paradigms (e.g. eye tracking movement or self-paced reading times). Despite the controversial status of formulaic language in the broadest sense in terms of its holistic representation, the role that recurrent word combinations play in language processing theory is not undermined. For future psycholinguistic studies which intend to use corpus-derived clusters as instruments, the relationship between different kinds of word combinations and the corresponding processing advantage will need to be further clarified. It is recommended to take into account the extent of formulaicity in those word combinations and possibly categorise them on the basis of degree of perceptual salience and

frequency before they can be used as experimental materials.

8.3.3 Second Language Teaching & Learning

As mentioned earlier, processing advantage and community recognition are possibly the major motivations that drive us to use formulaic language as opposed to creative language. In the context of writing, processing advantage does not seem as prominent as community recognition, at least on the part of writers when compared with the tremendous amount of information processing load from the real-time mode of spontaneous speech. For L2 learner writers, this could be a great relief as they can seek help from dictionaries or any other tools available when producing a piece of writing in a non-test context. To be more specific, for academic writing, the appropriate use of academic clusters enables L2 learners to construct an appropriate identity in the academic discourse community (Swales, 1990). The administration of a proper authorial voice, for instance, is one important aspect of this identity. As discussed in Sections 7.1.2.2 and 7.2.5, although the extent to which the authorial voice is explicitly presented is not found to directly affect writing quality, L2 learners can still benefit from the results of comparative analyses and learn how to establish their authorial voice in a similar way as native academics. ELT materials, therefore, serve as important sources for learners, especially when we have seen that learners can easily overuse the expressions introduced in ESL/EFL instruction (the phenomenon of 'lexical teddy bear' as termed by Hasselgreen, 1994). It is also possible to incorporate those native academic/literate clusters into computer-assisted writing software, which allows novice or learner writers to choose an appropriate expression from a pop-up menu which categorises word combinations on the basis of their structures or discourse functions. It is even possible for computers to automatically complete the most commonly used three-word, four-word, or longer word combinations when writers start the first part of an expression, which can effectively prevent novice/learner writers from using the deviant collocates in one formulaic sequence (e.g. *in recent years* instead of *in the*

recent years as discussed in Section 5.4.4.1).

Another feasible practice is to incorporate the stylistic and pragmatic recommendations discussed in Section 7.2 into style manuals for academic writing or dictionaries for advanced learners. As mentioned in Section 7.3, the 'Improving Your Writing Skills' section in the *Macmillan Dictionary for Advanced Learners* (Rundell, 2007) is an exemplar case for integrating English learning with learner-corpus research. Yet many of the findings in this thesis which can contribute to an effective formal writing style, e.g. the use of referential bundles with *-of*, are not covered by this pioneering dictionary.

A few studies, in fact, have reported the application of lexical bundles into language teaching. Cortes (2006) taught lexical bundles in a writing intensive history class, and Jones & Haywood (2004) also tried to gauge how effective teaching and learning lexical bundles could be. Although the students reported in these studies were not found to have used the instructed bundles to a substantial extent in writing, their awareness of formulaic sequences had increased. Since incorporating lexical bundles into the classroom is still in the earliest stage of development, more research has to be explored in this direction. In vocabulary studies, direct learning activities are usually recommended. Nation (2001) summarised the psycholinguistic process that deals with vocabulary learning as three phrases: 'noticing', 'retrieval' and 'generation'. In other words, learners internalise the vocabulary by firstly noticing and filling gaps, then retrieving the words from memory at the next encounter, and finally extending the knowledge about vocabulary when the retrieved words are found to be used differently with the previous encounter(s). The above conditions for vocabulary memorisation are considered the most efficient learning strategy (Takač, 2008, p. 75). If phraseology is treated in an identical way to vocabulary,[52] then the psycholinguistic principles generally adopted for vocabulary learning should be applicable to lexical bundles

[52] Part of phraseology can even be regarded as vocabulary if they are defined as holistic items.

as well (Coxhead, 2008, p. 155). The contribution from this thesis, accordingly, would be the results from a comparative perspective, which pinpoint the aspects which learner writing is most deviant from native writing and how it progresses across proficiencies. The two productive p-frames, '*the* + Noun + *of the/a*' and '*in the* + Noun + *of*' (cf. Section 5.4.4.1), are good examples which can be applied into language teaching to enhance learners' 'noticing' of the fundamental feature which characterises native academic writing.

The role of lexical bundles has been overlooked in language teaching and learning in the past, probably due to the fact that most lexical bundles lack perceptual salience. Yet we have seen that native writers can take advantage of expressions such as *in the context of* or *the extent to which* to add an academic/formal tone in writing. The native academic bundles, of course, require further editing and selection for pedagogical purposes. For instance, Ellis (2009) made use of a survey with language teachers in which *n-grams* (the same notion of *lexical bundles*, cf. Sections 3.1 & 3.2) were classified by means of mutual information (MI) to determine the best range of n-grams for ELT purposes. The issue as to how we can further select and incorporate the most appropriate bundles for language teaching and learning is certainly one new direction for future research.

8.3.4 SLA & Language Testing

In second language research, there has always been a compromise between data size and data quality. In the case of learner corpus studies, parsing and annotating L2 data is even more notoriously time- and manpower-consuming. Yet this thesis has provided a new approach in which a small proportion of highly frequent data can prove to be effective in describing the phraseological aspect of learner language, even without laborious error tagging and parsing.

With respect to the contribution towards language testing, this thesis is hoped to contribute to our understanding of learners' distinctive performance at various proficiency levels by shedding light on the discourse aspect of writing development. As we have seen,

learner proficiency can be determined by a set of robust rating procedures generally adopted in large-scale language proficiency tests. Such an integration serves as the interface between second language acquisition and language testing research as proposed by Bachman and Cohen (1998), with the incorporation of a new corpus and phraseological perspective. The findings in this thesis not only consolidate and complement existing research on second language development but can also be used to provide empirical underpinnings for a rating scale. The majority of existing rating scales, to my knowledge, are constructed on the basis of teachers' or researchers' perceptions about typical performance at defined levels rather than being drawn upon learners' actual performance. The CEFR referenced in this thesis has hence provoked some criticism about its lack of thorough empirical validation, particularly from the evidence obtained via learner data (e.g. Alderson, 2007; Hulstijn, 2007).

The notion of formulaicity and the stylistic features disclosed in this thesis have been discussed in comparison to the descriptors of the CEFR rating scale in Section 7.3. These core learner language features are seldom mentioned in the CEFR scale, yet the evidence provided from this thesis suggests that there exists distinctive pragmatic and stylistic development across proficiencies. As most current rating scales generally include lexis, grammar, and cohesion and coherence as the major definitive criteria for rating, it could be possible to consider adding discourse features as one of the criteria. At the same time, the results from quantitative and qualitative analyses might also contribute to development of an automated rating system.

8.4 Future Research Directions

As addressed in Section 1.3, this thesis intends to answer research questions from methodological, analytical, and explanatory aspects from the framework. In the course of searching for answers to the research questions, it is felt that this thesis is simply a start and far more research can be conducted in the future to better inform the areas of corpus

linguistics, phraseology, SLA, and language testing from various perspectives.

First of all, in terms of learner writing to be investigated, more learner data from different proficiency levels should be included in order to present a broader overview of writing development. It has to be noted that this thesis did intend to include learner writing from levels more than just CEFR-B2 and CEFR-C1. Yet it was found that spanning cross lower proficiency levels means that learner writing becomes increasingly less comparable, as far fewer argumentative and expository essays were produced at CEFR-B1. Additionally, in a preliminary analysis of the CEFR-B1 subcorpus with merely 26,356 words, the number of lexical bundles is also too small to be analysed. With a looser cut-off frequency and distribution set at three times or more in at least three texts, only 29 lexical bundles were generated. With an even lower threshold, three times in at least two texts, the number increases to 41 lexical bundles. As shown in Table 8-1, it can be seen that the nature of those CEFR-B1 bundles differs markedly from more advanced writing in that they look far more colloquial and personally involved. Interestingly, five frequent clusters, *on the other hand*, *at the same time*, *for a long time*, *is one of the*, and *all over the world* immediately come to light because they are also found in other groups of learner writing and overused by learners across various levels. Moreover, the expression *on the other hand* has been persistently ranked as the most frequently used bundle in all the learner groups investigated thus far. These learner-specific clusters, thus, appear to be fairly typical in learner writing regardless of proficiency levels.

Table 8-1 Lexical bundles in CEFR-B1 writing (bundle criteria altered to 3 times in 2 texts)[53]

Lexical bundles	Freq.	Dist.	Lexical bundles	Freq.	Dist.
on the other hand*	10	10	if you don't know	3	3
I think it is	8	8	it is because the	3	3
a lot of people	8	7	it is very important	3	3
have a lot of	7	7	that it is more	3	3
at the same time	6	6	the reason is that	3	3
if you want to	8	5	there are many people	3	3
are a lot of	4	4	think it is very	3	3
for a long time	4	4	I think this is	3	3
I hope I can	4	4	pay more for their	5	2
I would like to	4	4	***all over the world***	4	2
is one of my	4	4	for everyone to learn	4	2
more and more people	4	4	you want to do	4	2
there are a lot	4	4	have more time to	3	2
there are so many	4	4	I will feel very	3	2
there will be a	4	4	if you live in	3	2
is very important for	5	3	of having a child	3	2
is one of the	4	3	the people who live in	3	2
with a lot of	4	3	to go to work	3	2
are more and more	3	3	want to have a	3	2
become more and more	3	3	you will find the	3	2
I think the most	3	3			

*The bold italic font indicates the five learner-specific bundles.

For future research, various text-related variables such as task types in EAP-like writing should also be better controlled for a higher degree of comparability. For learner writing with specified proficiency levels, authentic candidate scripts produced in the context of language tests should be the priority data. For learner academic writing produced in the context of tertiary education, discipline-specific studies should be conducted, as recent research has suggested the existence of disciplinary variation in the use of lexical bundles (Hyland, 2008b). In this thesis, however, just as most second language studies, the accessibility of large

[53] These CEFR-B1 bundles have only been preliminarily refined. In other words, there might still be some context-dependent or overlapping bundles in this table.

amounts of learner data has always been a great challenge. With the use of the same BAWE-CH data, for instance, if we only focus on the data extracted from specified disciplines, the corresponding subcorpora would only get smaller. The balance between quantity and quality of data, therefore, might be worth more consideration in the future, particularly when corpus size appears to be an important factor in retrieving lexical bundles as well as frequency and dispersion thresholds.

In addition, it would be interesting to see whether learners from different L1 backgrounds would demonstrate a similar developmental pattern in terms of the use of lexical bundles. It is also possible to compare whether learners of a target language other than English would exhibit the same learner idiosyncrasies such as the tendency of making overstating claims or overreliance of VP-based bundles and discourse organisers. The last possibility with learner data is to look into the *atypical* learner performance at the specified levels (e.g. those with *misfit* values in Multi-faceted Rasch analysis indicative of rater disagreement as discussed in Section 6.1.2.2). Most of the second language developmental studies, to my knowledge, generally seek to identify learner performance characteristic of a specified proficiency level. From the perspective of language testing, however, atypical learner performance which often results in rater disagreement is equally important. If the problematic aspects which usually cause the raters to disagree with each other can be identified, we may be able to better understand learner language from an even more thorough perspective and a greater extent of rater consistency may be achievable.

In terms of the part of methodology, word combinations of various lengths, not just 4-word combinations, could be included so as to broaden the scope of defined frequency-driven formulaic language. Longitudinal studies can also be considered as they provide insight into use of formulaic language across writing development from a different perspective. As far as phraseology is concerned, it might be worth trying to apply the traditional top-down approach

with a pre-existing formulae list to contrast the frequency approach adopted in this thesis. In fact, the distinct features across proficiencies defined in Section 7.2 can be explored in this manner, which might throw more light on the discourse aspect and the phraseological use of learner language in each stage.

To sum up, there is still a tremendous amount of research to be completed in a variety of aspects involved in the current study. Although the purpose of exploring the use of lexical bundles in learner language development eventually would be aiming to facilitate the acquisition of these phraseological units, however, this is still very much uncharted territory. As Coxhead (2008, p. 158) points out, how can we realise phraseology research into practice of language teaching and learning 'when the nature of extent of these [phraseological] items has yet to be described?'

8.5 Concluding Remarks

The emergence of corpus linguistics has revolutionised the way that language is analysed. The comparative studies in this thesis shed light on aspects of language development which could not have been uncovered in the past. The tools which facilitate a corpus-driven approach make it possible to reveal patterns that linguists may have otherwise missed. The corpus-derived word combinations are found to generally lie across the boundary between grammar and lexis, functioning as the lexico-grammatical underpinnings of a language. Surely, frequency is not the only definitive indicator of formulaicity, and it is not the intention of this thesis to claim so. The corpus-driven frequency approach, however, does provide us with a new dimension to define formulaic language which stems from empirical evidence.

Moreover, the importance of corpus-extracted word combinations has been increasingly recognised in the way that discourse is structured. Despite this, the growing interest in identifying phraseology with corpus tools during the past decade does not appear to have encouraged ELT publishers or practitioners to put more emphasis on formulaic

language in the curriculum and/or materials. In the current thesis, through a review of lexical-bundle approach and investigation of three groups of academic writing and two CEFR-defined learner subcorpora, it was found that recurrent phraseology plays an important role in distinguishing language use across proficiencies, which has great implications for second language learning. It is argued that such frequency-driven formulaic expressions can be of help for learner writers to achieve native-likeness and hence should be integrated into ESL/EFL curricula after further research on the notion of lexical bundles. At the same time, the findings in this thesis also have implications for methodological issues in defining lexical bundles as well as other areas such as psycholinguistics and language testing.

Appendix 1 Peculiar Conditions of Overlapping Bundles

This part of document which deals with peculiar overlapping bundles supplements three major conditions of bundle overlaps, a) Complete Overlaps, b) Complete Subsumption, and c) Partial Subsumption, described in Section 3.6.2.

d) Peculiar Conditions: A number of overlapping bundles which do not fit any of the three major conditions of overlaps will be further discussed here. Due to some complex condition, in the following explicatory description of the system, X1, X2, (X3)... stand for the overlapping lexical bundles under examination while Y (or Y1, Y2) refers to the longer unit shared, which is placed below the dotted line in each row (i.e. each case of overlapping bundles). Again, a pair of brackets with the mark + was added in each finalised bundle to indicate the extended part of the longer unit.

There are four cases of peculiar overlapping bundles in Modular Study 1, all found in BAWE-EN and presented in Table 0-1. The first one is a pair of overlapping bundles (*be seen as a*, *can be seen as*), each of which can sustain even after deducting the occurrences of the longer shared unit Y (*can be seen as a*). In such a case, both of X1 and X2 would be retained, but the frequency in each bundle was deducted by one overlapping occurrence. A '+' mark was added preceding or following the bundle to indicate the position of overlaps.

The second peculiar case involves three lexical bundles, in which two of them (X1 and X2) demonstrate a complete one-to-one match in terms of occurrences and actually can combine into a longer expressions *it could be argued that*. The third bundle *be argued that the*, nevertheless, has only part of the occurrences identical with *it could be argued that* and therefore will be incorporated into the longer expression indicated in brackets. The frequency counts, according to Condition c), will be the frequency counts of X3 added to that of the shared longer unit Y1 deducting the repeated inclusion of Y2.

The third case also concerns three lexical bundles overlapping with one another. The

occurrences of X2 (*can be argued that*) are completely subsumed by X1 (*it can be argued*), which corresponds to Condition b) and should be represented as *it can be argued+(that)*. However, there is a third bundle X3 (*be argued that the*) with only one occurrence matched with *it can be argued that*. Remember that X3 *be argued that the* has already been discussed in the previous case *it could be argued that+(the)*, and as a matter of fact, the majority of occurrences of *be argued that the* are shared with this longer expression. Therefore, it is considered more sensible not to include *be argued that the* in the instance *it can be argued+(that)* here.

The last one is peculiar in the sense that the status of overlapping bundles satisfies Condition c), but the overlapped element, a preposition *for*, from X2 (*for the use* of) is not the majority of instances that occur prior to more frequent bundle X1 (*the use of the*). In the total eleven occurrences of *the use of the*, five of them follow a preposition while the other six do not. If X1 is combined with X2 as *(for)+the use of the*, then the structural categorisation of 'PP+of' would not be able to reflect the fact that over half of the occurrences do not follow a preposition. Therefore, it is decided to keep both X1 and X2 with the frequency of X1 being deducted with the overlapped two occurrences. Again, a '+' mark was added to indicate the position of overlaps.

Table 0-1 Peculiar overlapping bundles in Modular Study 1

Modular Study 1				
	Overlapping bundles	**Freq**	**Finalised bundles**	**Freq**
BAWE-EN	1. X1: *be seen as a*	6	*+be seen as a*	5
	X2: *can be seen as*	6	*can be seen as+*	5
	Y (X1+X2): *can be seen as a*	2		
	2. X1: *it could be argued*	12	*it could be argued that+(the)*	14
	X2: *could be argued that*	12		
	X3: *be argued that the*	7		
	Y1 (X1+X2): *it could be argued that*	12		
	Y2 (X1+X2+X3): *it could be argued that the*	5		
	3. X1: *it can be argued*	5	*it can be argued+(that)*	5
	X2: *can be argued that*	4		
	X3: *be argued that the*	7		
	Y1 (X1+X2): *it can be argued that*	4		
	Y2 (X1+X2+X3): *it can be argued that the*	1		
	4. X1: *the use of the*	11	*+the use of the*	9
	X2: *for the use of*	4	*for the use of+*	4
	Y (X1+X2): *for the use of the*	2		

In Modular Study 2, four cases of peculiar overlapping bundles were found in CEFR-B2 writing, and only one case ended up with a combined longer unit. As can be seen from Table 0-2, the first case, i.e. the combined bundle, originates from a trio of overlapping bundles, a mixed instance of complete subsumption and partial subsumption. The concordance lines of X2 (*are quite a lot*) are completely subsumed by X1 (*there are quite a*), but the overlapped six-word unit *there are quite a lot of* constitutes four out of five occurrences of X3 (*quite a lot of*), not all of them. Given that the longer overlapped unit is the majority of the three bundles, only the extended combination *there are quite a+(lot of)* would be kept in the finalised set. There are a number of other cases with slight overlapping. Sometimes one lexical bundle overlaps with several other word combinations, such as #3. Sometimes it involves a lexical bundle which has combined two bundles as having discussed earlier, such as *there are quite a+(lot of)* in #2. No change would be made in these instances because the frequency of the overlapped unit is slim, usually with just one occurrence. These

individual lexical bundles would therefore all be kept. Similarly, a '+' mark was added in each case to indicate the position of overlaps.

Table 0-2 Peculiar overlapping bundles in Modular Study 2

Modular Study 2				
	Overlapping bundles	**Freq**	**Finalised bundles**	**Freq**
CEFR-B2	1. X1: *there are quite a*	5	*there are quite a+(lot of)*	5
	X2: *are quite a lot*	4		
	X3: *quite a lot of*	5		
	Y (X1+X2+X3): *there are quite a lot of*	4		
	2. X1: *a lot of problem(s)*	16	*+a lot of problem(s)*	16
	X2: *there are quite+(lot of)*	5	*there are quite a+(lot of)+*	5
	X3: *has a lot of*	4	*has a lot of+*	4
	Y1 (X1+X2): *there are quite a lot of problem(s)*	1		
	Y2 (X1+X3): *has a lot of problem(s)*	1		
	3. X1: *a lot of problem(s)*	16	*+a lot of problem(s)*	16
	X2: *there are a lot of*	11	*there are a lot of+*	11
	X3: *has a lot of*	4	*has a lot of+*	4
	X4: *have a lot of*	4	*have a lot of+*	4
	Y1 (X1+X2): *there are a lot of problem(s)*	1		
	Y2 (X1+X3): *has a lot of problem(s)*	1		
	Y3 (X1+X4):*have a lot of problem(s)*	1		
	4. X1: *as the result of*	4	*as the result of+*	4
	X2: *the result of the*	4	*+the result of the*	4
	X3: *the result of this*	5	*+the result of this*	5
	Y1 (X1+X2): *as the result of the*	1		
	Y2 (X1+X3): *as the result of this*	1		

Appendix 2 Common Reference Levels: Global Scale

(Common European Framework of Reference, Council of Europe, 2001, p. 24)

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Appendix 3 CEFR Written Assessment Criteria Grid

(Relating Language Examinations to the Common European Framework of Reference for Languages, Council of Europe, 2003, p. 187)

	Overall	Range	Coherence	Accuracy	Description	Argument
C2	Can write clear, highly accurate and smoothly flowing complex texts in an appropriate and effective personal style conveying finer shades of meaning. Can use a logical structure which helps the reader to find significant points.	Shows great flexibility in formulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent and highly accurate grammatical control of even the most complex language forms. Errors are rare and concern rarely used forms.	Can write clear, smoothly flowing and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.	Can produce clear, smoothly flowing, complex reports, articles and essays which present a case or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured and mostly accurate texts of complex subjects. Can underline the relevant salient issues, expand and support points of view at some length with subsidiary points, reasons and relevant examples, and round off with an appropriate conclusion.	Has a good command of a road range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. The flexibility in style and tone is somewhat limited.	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; occasional errors in grammar, collocations and idioms.	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in a mostly assured, personal, natural style appropriate to the reader in mind.	Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support point of view with some subsidiary points, reasons and examples.
B2	Can write clear, detailed official and semi-official texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. Can make a distinction between formal and informal language with occasional less appropriate expressions.	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic.	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading.	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more common situations. Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.	Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on a familiar routine and non-routine matters, within his field with some confidence. Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
A2	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". Longer texts may contain expressions and show coherence problems which make the text hard to understand.	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. Errors may sometimes cause misunderstandings.	Can write very short, basic descriptions of events, past activities and personal experiences. Can write short simple imaginary biographies and simple poems about people.	
A1	Can write simple isolated phrases and sentences. Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand.	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. Errors may cause misunderstandings.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.	

Appendix 4 Lexical Bundles in Frequency Order

CEFR-B2		CEFR-C1		BAWE-CH		BAWE-EN		FLOB-J	
is one of the	18	on the other hand	28	on the other hand	36	in the case of	23	in the context of+(the/a)	19
on the other hand	18	at the same time	14	(and)+at the same time	24	can be used to	19	in the case of	19
at the same time	17	is one of the	14	that there is a/an	17	as a result of+(the)	17	on the other hand	19
+a lot of problem(s)	16	I would like to	11	as well as the	16	it is important to	17	the nature of the	17
(can)+have the right to	14	it is obvious that+(the)	11	in the long run	13	it could be argued that+(the)	14	as a function of	15
for a long time	11	one of the most	11	as a result of	12	(at)+the end of the	13	on the basis of	14
there are a lot of+	11	as well as the	7	one of the most	11	in terms of the	13	in terms of the	14
a lot of people	10	it is believed that	7	can be used to	10	(for)+the development of the	13	it is necessary to	14
a lot of time	9	for a long time	6	in the case of	10	it is possible to	13	the way in which+(the)	14
I would like to	7	in the process of	6	is one of the	9	is one of the	12	(at)+the end of the	13
it is also a	7	it is true that	6	it is difficult to	9	the rest of the	12	it is clear that	11
most of them are	7	the end of the	6	as one of the	8	it can be seen+(that)	12	the rest of the	11
the most important thing+(is)	7	the quality of the	6	at the end of+(the)	8	one of the main	11	one of the most	10
and a lot of	6	the rest of the	6	in order to achieve	8	as well as the	10	at the same time	10
become more and more	6	we can see that	6	it is necessary to	8	+the use of the	9	by the fact that	9
it is difficult to	6	a great deal of	5	the rest of the	8	in the same way	9	a wide range of	9
it is very difficult+(to)	6	all over the world	5	can be explained by	7	to be able to	8	as a result of	9
will not be able to	6	as a matter of+(fact)	5	in order to avoid	7	(due)+to the fact that	8	that there is a	9
with the development of	6	at the beginning of+(the)	5	in the end of+(this)	7	an example of this+(is)	8	per cent of the	9
(from)+my point of view	5	in order to make	5	we can see that	7	in the form of	8	in the form of	8
(is)+the best way to	5	in such a way+(that)	5	a number of factors	6	it is clear that	8	on the one hand	8
a great number of	5	is a kind of	5	all over the world	6	the fact that the	8	the extent to which	8
all over the world	5	we can say that	5	as long as the	6	in order to make	8	in the sense that	8
has the right to	5	as a result of	4	at the beginning of	6	the nature of the	7	would have to be	8
is based on the	5	as far as the	4	at the expense of	6	one of the most+(important)	7	as we have seen	8
is very important to	5	can be divided into	4	in the recent years	6	and as a result	7	in the presence of	8

most of the people	5	has the right to	4	it has been suggested that	6	it is necessary to	7	in the absence of	7
one of the most	5	how to deal with	4	it is easy to	6	the way in which	7	is likely to be	7
people who live in	5	it is hard to	4	that is to say	6	can be applied to+(the)	7	the size of the	7
the main reason is	5	it is not easy+(for)	4	to the development of	6	than that of the	7	are more likely to	7
+the result of this	5	it is very difficult	4	(played)+an important role in	5	and the use of	6	as we shall see	7
there are quite a+(lot of)+	5	necessary for us to	4	are more likely to	5	are more likely to	6	that there is no	7
there are too many	5	on the basis of	4	as a part of	5	not be able to	6	the history of the	7
want to be a	5	some of them are	4	as part of a	5	the extent to which	6	the turn of the century	7
a large amount of	4	the relationship between the	4	bear in mind that	5	was one of the	6	can be found in	6
a very important role	4	there are still some	4	in order to be	5	could be used to	6	in the first place	6
all of them are	4	to cope with the	4	in terms of the	5	is an example of+(a)	6	it is difficult to	6
and to be a	4			in the context of	5	it is difficult to	6	it is possible to	6
are not allowed to	4			it is believed that	5	the structure of the	6	of a number of	6
as a matter of fact	4			last but not least	5	can be found in	6	on the part of	6
as I have mentioned	4			on the basis of	5	is the fact that	6	in so far as	6
as the result of+	4			the nature of the	5	the length of the	6	in the number of	6
as we all know	4			a high level of	4	with respect to the	6	it is important to	6
because they are not	4			a large number of	4	would have to be	6	seems to have been	6
bring a lot of	4			a wide range of	4	can be seen as+	5	with respect to the	6
but there are still	4			as part of the	4	+be seen as a	5	are shown in fig	6
has a lot of+	4			as soon as the	4	at the same time	5	can be used to	6
have a lot of+	4			can be divided into	4	be taken into account	5	in the light of	6
him or her to	4			can be regarded as	4	be used in the	5	a function of time	6
I am going to	4			essay is going to	4	could be seen as	5	as shown in fig	6
I think it is	4			for the development of	4	it would have been	5	be taken into account	5
I think that this	4			in addition to the	4	this is due to+(the)	5	for each of the	5
if there is a	4			in order to maintain	4	with the development of	5	in a number of	5
in the following paragraphs	4			in order to make	4	would be able to	5	it can be seen that	5
is more important than	4			in order to understand	4	are likely to be	5	to the fact that	5
is the most important	4			in the form of	4	can also be used+(to)	5	would be difficult to	5
is totally different from	4			is considered to be	4	in relation to the	5	at the time of	5
it is a good	4			is illustrated in figure	4	in terms of its	5	be found in the	5

it is a very	4			is not only a	4	it can be argued+(that)	5	in the hands of	5
it is not a	4			it can be seen	4	of the number of	5	the degree to which	5
it is true that	4			it has to be	4	through the use of	5	the role of the	5
should learn how to	4			it is easy for	4	to a lack of	5	the rules of the	5
some of them are	4			it is important to	4	can be used for	5	the strength of the	5
some people think that+(the)	4			meet the requirement of	4	for each of the	5	the value of the	5
the end of the	4			must be able to	4	there would be no	5	to be able to	5
the quality of the	4			of the number of	4	a great deal of	4	a function of the	5
the rest of the world	4			pay more attention to	4	an integral part of	4	in the course of	5
+the result of the	4			the development of the	4	as part of the	4	there is evidence that	5
there are so many	4			the importance of the	4	by the presence of	4	whether or not to	5
there will be a	4			the role of the	4	in an attempt to	4	a large number of	4
we can see the	4			the size of the	4	is by no means	4	an example of this	4
				the top of the	4	it is estimated that	4	be seen in the	4
				this is due to	4	it should be noted	4	by a variety of	4
				this means that the	4	of some of the	4	in contrast to the	4
				to be able to	4	on the other hand	4	in more detail in	4
				to ensure that the	4	should be able to	4	in relation to the	4
				to the fact that	4	the fact that they	4	in terms of a	4
				will focus on the	4	there is no evidence	4	in view of the	4
				with respect to the	4	this means that the	4	it has been suggested	4
				with the introduction of	4	to a certain extent	4	it has not been	4
						to be added to	4	it is not always	4
						to enable them to	4	on a number of	4
						to take into account	4	the fact that this	4
						would need to be	4	the right hand side	4
						as a way of	4	the status of the	4
						at the heart of	4	the structure of the	4
						be included in the	4	the ways in which	4
						because it is not	4	to a large extent	4
						can be seen in	4	a high level of	4
						for the use of+	4	are likely to be	4

						in order to minimise	4	at each end of	4
						in the absence of	4	at the beginning of	4
						in this essay I	4	end of the spectrum	4
						should be placed on	4	has a number of	4
						taking into account the	4	in the face of	4
						that is to say	4	is concerned with the	4
						that need to be	4	it is not clear	4
						the quality of the	4	the creation of a	4
						the size of the	4	the existence of a	4
						this may be due to	4	the impact of the	4
						to cope with the	4	the magnitude of the	4
						will be able to	4	the point of view	4
						will be used to	4	the results of the	4
						with the addition of	4	the second half of	4
								to that of the	4
								was followed by a	4
								was not so much	4
								was one of the	4

References

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language* Teaching (pp. 55-76). Amsterdam/Philadelphia: John Benjamins.

Alderson, C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, 91, 659-663.

Alderson, C. (1990). Bands and scores. In C. Alderson & B. North (Eds.), *Language Testing in the 1990s* (pp. 71-86). London: Modern English Publications and the British Council.

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1), 71-83.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 101-122). Oxford: Oxford University Press.

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

Aston, G., Bernardini, S., & Stewart, D. (Eds.). (2004). *Corpora and language learners*. Amsterdam; Philadelphia: John Benjamins.

Atkins, S., & Clear, J. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1-16.

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge, U.K.; New York, NY: Cambridge University Press.

Bailey, S. (2006). *Academic writing: a handbook for international students* (2nd ed.). London: Routledge.

Ball, F. (2002). Developing wordlists for BEC. Cambridge ESOL: *Research Notes* (8), 10-13.

Ball, F. (2001). Using corpora in language testing. Cambridge ESOL: *Research Notes* (6), 6-8.

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels (IELTS Funded Research Project, Round 10, 2004).

Bennett, K. (2009). English academic style manuals: A survey. *Journal of English for Academic Purposes*, 8, 43-54.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, 263-286.

Biber, D. (2006). *University language : a corpus-based study of spoken and written registers*. Amsterdam ; Philadelphia, Penn.: J. Benjamins.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at ...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. Princeton, NJ: Educational Testing Service.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* (pp. 71-92). Frankfurt: Peter Lang.

Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversations and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: studies in honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

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

Brace, N., Kemp, R., Snelgar, R. (2003). SPSS for Psychologists. Basingstoke and New York: Palgrave Macmillan.

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Boyle, A., & Booth, D. (2000). The UCLES/CUP learner corpus. Cambridge ESOL: *Research Notes* (1), 10.

Cambridge ESOL (2009). IELTS candidate performance 2008. Cambridge ESOL: *Research Notes* (36), 30-32.

Cambridge ESOL (2007). Sample papers for the Cambridge ESOL General English suite of exams, from http://www.cambridgeesol.org/exams/index.html

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing*, 3, 179-201.

Chen, C. W.-y. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, 11(1), 113-131.

Chuang, F.-Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora*, 1(2), 251-271.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), 72-89.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17, 391-406.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23, 397-423.

Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Cosme, C. (2006). Clause combining across languages. A corpus-based study of English-French translation shifts. *Languages in Contrast*, 6(1), 71-108.

Council of Europe (2004). *Reference supplement to the preliminary pilot version of the Manual for Relating Language Examinations to the CEFR*. Strasbourg: Language Policy Division.

Council of Europe (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF). Manual: Preliminary Pilot Version*. DGIV/EDU/LANG 2003, 5. Strasbourg: Language Policy Division.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 1-20). Oxford: Oxford University Press.

Cowie, A. P. (Ed.). (1998b). *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, 2(3), 223-235.

Coxhead, A. (2008). Phraseology and English for academic purposes. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 149-161). Amsterdam: John Benjamins.

Crompton, P. (1997). Hedging in academic writing: some theoretical problems. *English for Specific Purposes*, 16(4), 271-287.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, 26(2), 163-174.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

De Cock, S. (2007). Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight. In M. C. Campoy & M. J. Luzon (Eds.), *Spoken Corpora in Applied Linguistics* (pp. 217-233). Bern: Peter Lang.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgium Journal of English and Literatures (BELL)*, New Series 2: 225-246.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 51-68). Amsterdam: Rodopi.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59-80.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London: Longman.

Develle, S. (2008). The revised IELTS Pronunciation scale. Cambridge ESOL: *Research Notes*(34), 36-39.

Ellis, N. C. (2009). Central issues in Corpora and SLA – Analysing the input, the intake, and the reasons for their differences. Paper presented at the *ICAME*.

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

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29-62.

Field, Y., & Yip, L. M. O. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal*, 23(1), 15-28.

Fletcher, W. (2003-2006). Phrases in English. http://pie.usna.edu

Flowerdew, L. (2000). Investigating referential and pragmatic erorrs in a learner corpus. In L. Burnard & T. McEnery (Eds.), Rethinking language pedagogy from a corpus perspective: papers from the Third International Conference on Teaching and Language Corpora (pp. 145-154). Frankfurt: Peter Lang.

Gilquin, G., & Paquot, M. (2007). Spoken Features in Learner Academic Writing: Identification, Explanation and Solution. Paper presented at the Corpus Linguistics Conference (CL 2007), University of Birmingham, UK.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-335.

Gilquin, G. (2002). Automatic retrieval of syntactic structures. The quest for the Holy Grail. *International Journal of Corpus Linguistics*, 7(2), 183-214.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam & Philadelphia: John Benjamins.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & Meunier (Eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2005). Pushing back the limits of phraseology. How far can we go? In C. Cosme, Gouverneur, C., Meunier, F. & Paquot, M. (Ed.), *Proceedings of the Phraseology 2005 Conference* (pp. 165-168). Université catholique de Louvain: Louvain-la-Neuve.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO* Journal, 20(3), 465-480.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (Vol. 6, pp. 3-33). Amsterdam/Philadelphia: John Benjamins.

Granger, S. (1998a). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner Language on Computer* (pp. 3-18). London and New York: Addison Wesley Longman.

Granger, S. (1998b). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 145-160). Oxford: Oxford University Press.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on Computer* (pp. 119-131). Londond and New York: Addison Weslesy Longman.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1), 17-27.

Granger, S. (1993a). The International Corpus of Learner English. *The European English Messenger*, 2(1), 34.

Granger, S. (1993b). The International Corpus of Learner English. In J. Aarts, de Haan, P. and Oostdijk, N (Ed.), *English Language Corpora: Design, Analysis and Exploitation* (pp. 57-69). Amsterdam: Rodopi.

Gries, S. T. (2008). Phraseology and linguistic theory. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3-25). Amsterdam & Philadelphia: John Benjamins.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics : a study of particle placement*. New York: Continuum.

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

Hasselgreen, A. (1994). Lexical Teddy Bears and Advanced learners: a study into the ways Norwegian students cope with vocabulary. *International Journal of Applied Linguistics*, 4(2), 237-260.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9(2), 122-159.

Hawkey, R. (2001). Towards a common scale to describe L2 writing. Cambridge ESOL: *Research Notes*(5), 9-13.

Haywood, S., & Jones, M. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 269-291). Amsterdam: John Benjamins.

Helms-Park, R., & Stapleton, P. (2003). Questioning the importance of individualized voice in undergraduate L2 argumentative writing: An empirical study with pedagogical implications. *Journal of Second Language Writing*, 12, 245-265.

Howarth, P. (1998a). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24-44.

Howarth, P. (1998b). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 161-186). Oxford: Oxford University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and Qualitative dimensions of language proficiency. *The Modern Language Journal*, 91, 663-667.

Hundt, M., Sand, A., & Siemund, R. (1998). Manual of Information to accompany The Freiburg - LOB Corpus of British English ('FLOB') Available from http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University 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(1), 4-21.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*, 34, 1091-1112.

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. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13(3), 239-156.

James, C. (1992). Awareness, consciousness and language contrast. In C. Mair & M. Markus (Eds.), *New Departures in Contrastive Linguistics* (pp. 183-197). Innsbruck: Innsbrucker Beitrage zur Kulturwissenschaft.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91, 433-445.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 269-291). Amsterdam: John Benjamins.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.

Jordan, R. R. (1999). *Academic writing course: study skills in English* (3rd ed.). Harlow: Longman.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies* (pp. 106-129). Strasbourg, France Council of Europe.

Kennedy, C., & Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor & P. Falvey (Eds.), *IELTS collected paper: research in speaking and writing assessment* (pp. 316-378). Cambridge: Cambridge University Press.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187-217.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance *Language Testing*, 19(1), 1-31.

Leech, G. (1993). Corpus annotation schemes. *Lit Linguist Computing*, 8(4), 275-281.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of Second Language Writing*, 18, 85-102.

Linacre, J. M. (2008). *Facets Rasch measurement computer program* (Version 3.64.0). Chicago: Winsteps.com.

Lorenz, G. (1999). *Adjective intensification--Learners versus native speakers. A corpus study of argumentative writing*. Amsterdam: Radopi.

Lorenz, G. (1998). Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on Computer* (pp. 53-66). London and New York: Addison Wesley Longman Limited.

Lüdeling, A., Adolphs, P., Kroymann, E., & Walter, M. (2005). Multi-level error annotation in learner corpora. *Proceedings of the Corpus Linguistics 2005 Conference*, Birmingham, UK,14-17 July 2005.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12, 54-71.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1), 1-31.

Mayor, B., Hewings, A., North, S., Swann, J., & Coffin, C. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In L. Taylor & P. Falvey (Eds.), *IELTS collected paper: research in speaking and writing assessment* (pp. 250-315). Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics* (2 ed.). Edinburgh: Edinburgh University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London ; New York: Longman.

Meunier, F., & Granger, S. (Eds.). (2007). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: John Benjamins Publishing.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on Computer* (pp. 19-37). New York: Addison Wesley Longman.

Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Cambridge, UK ; New York: Cambridge University Press.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on Computer* (pp. 186-198). London and New York: Longman.

Milton, J., & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In L. Flowerdew & A. K. K. Tong (Eds.), *Entering Text* (pp. 127-143). Hong Kong: Language Centre, The Hong Kong University of Science and Technology.

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

Moon, R. (1998a). *Fixed expressions and idioms in English: a corpus-based approach*. Oxford and New York: Clarendon Press and Oxford University Press.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 79-100). Oxford: Oxford University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge Univeristy Press.

Nicholls, D. (2003). The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson & A. McEnery (Eds.), *Proceedings of Corpus Linguistics* 2003 (pp. 571-582).

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217-263.

Oakey, D. (2009). Fixed Collocational Patterns in Isolexical and Isotextual Versions of a Corpus. In P. Baker (Ed.), *Contemporary Corpus Linguistics*. London: Continuum.

Oakey, D. (2002). Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 111-129). Amsterdam and Philadelphia: John Benjamins Publishing Company.

Oshima, A., & Hogue, A. (1999). *Writing Academic English* (3rd ed.). White Plains, NY: Longman.

Pendar, N., & Chapelle, C. (2008). Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal*, 25(2), 189-206.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on Computer* (pp. 107-118). London and New York: Addison Wesley Longman.

Poos, D., & Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 3-23). Amsterdam and Philadelphia: John Benjamins Pubishing Company.

Quirk, R., & Greenbaum, S. (1973). *A university grammar of English*. Harlow: Longman.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1972). *A Grammar of contemporary English*. London: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English. In S. Granger (Ed.), *Learner English on Computer* (pp. 41-52). London and New York: Addison Wesley Longman Limited.

Romer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 151-168). Amsterdam; Philadelphia: John Benjamins.

Rosen, L. J. (2008). *The academic writer's handbook* (2nd ed.). Harlow: Longman.

Rundell, M. (Ed.). (2007). *Macmillan English Dictionary for Advanced Learners* (Second Edition). Oxford: Macmillan Education.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25, 465-493.

Schmitt, N. (2004). *Formulaic sequences: acquisition, processing, and use*. Amsterdam; Philadelphia: John Benjamins.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127-151). Amsterdam: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127-152). Amsterdam: John Benjamins Publishing.

Scott, M. (2007). Oxford WordSmith Tools 4.0.

Selinker, L. (1972). Interlanguage. *IRAL*, 10(3), 209-231.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Boca Raton, Fla. ; London: Chapman & Hall/CRC.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Skelton, J. (1988). The care and maintenance of hedges. *ELT Journal*, 42(1), 37-43.

Strzalkowski, T. (Ed.). (1998). *Natural Language Information Retrieval*. Dordrecht: Kluwer.

Stubbs, M. (2007a). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 89-105). Amsterdam: Radopi.

Stubbs, M. (2007b). Quantitative data on multi-word sequences in English: The case of word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, Discourse and Corpora: Theory and Analysis*. London: Continuun.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7(2), 215-244.

Svensson, M. H. (2008). A very complex criterion of fixedness: Non-compositionality. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 81-93). Amsterdam; Philadelphia: John Benjamins.

Swales, J., & Feak, C. B. (1994). *Academic writing for graduate students : essential tasks and skills: a course for nonnative speakers of Englis*h. Ann Arbor: University of Michigan Press.

Swales, J. (1990). *Genre analysis : English in academic and research settings*. Cambridge England ; New York: Cambridge University Press.

Takač, V. P. (2008). *Vocabulary learning strategies and foreign language acquisition* Clevedon; Buffalo; Toronto: Multilingual Matters.

Teubert, W., & Krishnamurthy, R. (Eds.). (2007). *Corpus Linguistics*. London: Routledge.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44(2), 307-336.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam; Philadelphia: J. Benjamins.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 155-172). Amsterdam: John Benjamins.

Van Moere, A. (2006). Validity evidence in a universal group oral test. *Language Testing*, 23(4), 411-440.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity* (Vol. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32(4), 213-231.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: an integrated model. *Language & Communication*, 20, 1-28.

Wright, B. D., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

Zhao, C. G., & Llosa, L. (2008). Voice in high-stakes L1 academic writing assessment: Implications for L2 writing instruction. *Assessing Writing*, 13, 153-170.