A CORPUS-DRIVEN STUDY OF EXTENDED WRITING TOWARDS THE DEVELOPMENT OF A VOCABULARY AND PHRASEOLOGY INDEX

WONG WEI LUN

UNIVERSITI PENDIDIKAN SULTAN IDRIS

2024

A CORPUS-DRIVEN STUDY OF EXTENDED WRITING TOWARDS THE DEVELOPMENT OF A VOCABULARY AND PHRASEOLOGY INDEX

WONG WEI LUN

THESIS IS SUBMITTED IN FULFILMENT OF THE REQUIREMENT OF THE DOCTOR OF PHILOSOPHY

FACULTY OF LANGUAGES AND COMMUNICATION
SULTAN IDRIS EDUCATION UNIVERSITY

2024

UPSI/PS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS

Sila tanda (√)
Kertas Projek
Sarjana Penyelidikan
Sarjana Penyelidikan dan Kerja Kursus
Doktor Falsafah

INSTITUT PENGAJIAN SISWAZAH

PERAKUAN KEASLIAN PENULISAN

Perakuan ini telah dibuat pada 9 (hari bulan) Julai (bulan) 20..24.

i. Perakuan pelajar :

Saya, Wong Wei Lun, P20211000930, Faculty of Languages and Communication (SILA NYATAKAN NAMA PELAJAR, NO. MATRIK DAN FAKULTI) dengan ini mengaku bahawa disertasi/tesis yang bertajuk A Corpus-Driven Study of Extended Writing Towards the Development of a Vocabulary and Phraseology Index

adalah hasil kerja saya sendiri. Saya tidak memplagiat dan apa-apa penggunaan mana-mana hasil kerja yang mengandungi hak cipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hak cipta telah dinyatakan dengan sejelasnya dan secukupnya

Colin

Tandatangan pelajar

ii. Perakuan Penyelia:

Saya, Associate Professor Dr. Mazura Mastura Muhammad (NAMA PENYELIA) dengan ini mengesahkan bahawa hasil kerja pelajar yang bertajuk A Corpus-Driven Study of Extended Writing Towards the Development of a Vocabulary and Phraseology Index

(TAJUK) dihasilkan oleh pelajar seperti nama di atas, dan telah diserahkan kepada Institut Pengajian Siswazah bagi memenuhi sebahagian/sepenuhnya syarat untuk memperoleh Ijazah Doctor of Philosophy (TESL) (SILA NYATAKAN NAMA IJAZAH).

09/07/2024

Tarikh

Tandatangan Penyelia

UPSI/IPS-3/BO 31
Pind.: 01 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: A Corpus-Driven Study of Extended Writing Towards the Development of a Vocabulary and Phraseology Index

No. Matrik /*Matric's No.*: P20211000930

Saya / *I* : Wong Wei Lun

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah / *Please tick (√) for category below:-*

☐ **SULIT/*CONFIDENTIAL*** — Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*

☐ **TERHAD/*RESTRICTED*** — Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restricted information as specified by the organization where research was done.*

☑ **TIDAK TERHAD / *OPEN ACCESS***

Colin

(Tandatangan Pelajar/ Signature)

Tarikh: 09/07/2024

(Tandatangan Penyelia / *Signature of Supervisor*
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Assoc. Prof. Dr. Mazura@Mastura Binti Muhammad
Dean
Faculty of Languages and Communication
Sultan Idris Education University
35900 Tanjong Malim, Perak

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

ACKNOWLEDGEMENTS

Completing a PhD thesis is a challenging journey that requires dedication, persistence, and support from others. I am grateful to have been surrounded by an incredible community of individuals who have helped me navigate this journey and who have made it all the more fulfilling.

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Mazura, for her unwavering support, guidance, and mentorship throughout this process. Her expertise, patience, and insightful feedback have been invaluable in shaping this work.

I would also like to thank Dr. Charanjit for being a supportive scholar and for sharing her expertise and knowledge with me.

To my best friends and comrades CKR, Mairas, Dr. Warid, Wong Yoke Yee, Chris Chong, Tham Wei Wen, and Chai Yuen Yee, thank you for your long-term support, cheers, and for being a constant source of inspiration and encouragement. Your unwavering belief in me and my abilities have kept me motivated throughout this journey.

Finally, I would like to acknowledge the countless individuals who have contributed to this work in big and small ways. Your insights, feedback, and contributions have been invaluable from research participants to fellow academics.

To everyone who has played a role in this journey, I offer my deepest thanks. This thesis is a reflection not only of my own efforts but of the collective efforts of a community of individuals who have supported, encouraged, and challenged me throughout this process.

ABSTRACT

This study aimed to develop a vocabulary and phraseology index using a corpus-driven approach, addressing three key objectives, which are examining English vocabulary, phrases, and semantic domains in extended writing by upper primary learners across Malaysia; exploring contributing factors of salient vocabulary and phrase occurrences; and developing a vocabulary and phraseology index. A mixed-method approach and sequential explanatory design underpinned this study. Purposive sampling led to the selection of 28 primary schools, from which 560 extended writing samples (comprising 152,187 words) were gathered. Analytical tools like LancsBox 6.0, W-Matrix, USAS semantic tagger, and Berkeley Neural Parser were utilised to identify salient vocabulary, phrases, and semantic domains. The quantitative analysis highlighted both similarities and differences in vocabulary (you, friends, know, beautiful, when; this, video, decided, violent, together) and phrase (in the, a lot of, my family and I; the fox, on the island, I will be sharing) usage across regions, as well as in semantic domains (food; green issues). For qualitative data, it was determined that instructional guidance and learning resources were the key contributing factors to such occurrences. The process of developing the index involved compiling a learner corpus, categorising literacy genres, and aligning salient vocabulary and phrases with CEFR levels (A1-C2). The findings offer valuable insights for policymakers and educators in enhancing ESL teaching methodologies through a deeper understanding of learner language use. The findings implied that English teachers could clearly understand the salient vocabulary and phrases required for extended writing, and that learners are equally informed about the appropriate vocabulary and phrases to employ in such contexts. Furthermore, these insights may serve as a benchmark for policymakers seeking to refine language policies to bolster English as a Second Language (ESL) education in Malaysian primary schools.

SATU KAJIAN BERDASARKAN KORPUS TENTANG PENULISAN LANJUTAN UNTUK PEMBANGUNAN INDEKS KOSA KATA DAN FRASALOGI

ABSTRAK

Kajian ini bertujuan untuk mengembangkan indeks kosa kata dan frasalogi menggunakan pendekatan berdasarkan korpus, dengan menangani beberapa objektif utama: memeriksa kosa kata, frasa, dan domain semantik dalam penulisan lanjutan oleh pelajar sekolah rendah atas di seluruh Malaysia; menyelidiki faktor-faktor yang menyumbang kepada kejadian kosa kata dan frasa yang menonjol; serta mengembangkan indeks kosa kata dan frasalogi. Kajian ini menggunakan pendekatan metodologi campuran dan reka bentuk penjelasan berurutan. Sampel kajian ini melibatkan 28 buah sekolah rendah, iaitu 560 buah sampel penulisan lanjutan (terdiri daripada 152,187 perkataan) dikumpulkan. Alat analisis seperti LancsBox 6.0, W-Matrix, penanda semantik USAS, dan Berkeley Neural Parser digunakan untuk analisis kuantitatif kosa kata, frasa, dan domain semantik yang menonjol. Kajian ini juga menggunakan kaedah temu bual terhadap 28 orang guru Bahasa Inggeris untuk mengenal pasti faktor-faktor penyumbang. Dapatan kajian ini menunjukkan bahawa terdapat persamaan dan perbezaan dalam penggunaan kosa kata (anda, kawan, tahu, cantik, bila; ini, video, memutuskan, ganas, bersama) dan frasa (di, banyak, saya dan keluarga saya; serigala itu, di pulau, saya akan berkongsi) di seluruh wilayah, serta dalam domain semantik (makanan; isu hijau). Selain itu, terdapat pengaruh yang menyebabkan faktor-faktor penyumbang seperti bimbingan pengajaran dan sumber pembelajaran. Proses mengembangkan indeks melibatkan pengumpulan korpus pelajar, pengkategorian genre literasi, dan menyelaraskan kosa kata dan frasa yang menonjol dengan tahap CEFR (A1-C2). Penemuan ini mengimplikasikan bahawa guru-guru Bahasa Inggeris dapat memahami dengan jelas perbendaharaan kata dan frasa penting yang diperlukan untuk penulisan lanjutan, dan para pelajar juga mempunyai pengetahuan yang setara mengenai perbendaharaan kata dan frasa yang sesuai untuk digunakan dalam konteks tersebut. Selain itu, pandangan ini boleh dijadikan sebagai penanda aras bagi pembuat dasar yang berusaha menyempurnakan dasar bahasa untuk mengukuhkan pendidikan Bahasa Inggeris sebagai Bahasa Kedua (ESL) sekolah rendah di Malaysia.

CONTENTS

	Page
DECLARATION OF ORIGINAL WORK	ii
DECLARATION OF DISSERTATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
ABSTRAK	vi
CONTENTS	vii
LIST OF TABLES	xiv
LIST OF FIGURES	xviii
LIST OF ABBREVIATIONS	xxiv
LIST OF APPENDICES	xxvi

CHAPTER 1 INTRODUCTION

1.1	Introduction	1
1.2	Background of the Study	3
1.3	Problem Statement	8
1.4	Purpose of the Study	17
1.5	Objectives of the Study	17
1.6	Research Questions	18
1.7	Significance of the Study	18
1.8	Study Limitations	21

1.9 Operational Definition 22

1.9.1 Vocabulary 22

1.9.2 Phrases 23

1.9.3 Advanced Malaysian Upper Primary School Learners 23

1.9.4 Corpus-Driven Study 24

1.9.5 Extended Writing 25

1.9.6 Semantic Domains 27

1.9.7 Vocabulary and Phraseology Index 27

1.9.8 Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC) 29

1.10 Summary 29

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction 31

2.2 Corpus Linguistics (CL) 32

2.2.1 Definition of Corpus Linguistics 38

2.2.2 Corpus Linguistics in Malaysia 45

2.3 History of Corpus Linguistics 49

2.4 Types of Corpora 52

2.5 Phraseology 57

2.6 N-gram Theory and Models 59

2.6.1 Application of N-grams 62

2.6.2 Simple N-gram 63

2.7 Corpus-based versus Corpus-driven Approach 65

2.7.1 Corpus-based Approach (CBA) 66

2.7.2 Corpus-driven Approach (CDA) 67

2.7.3 Reconceptualising the Differences 68

2.7.4 Annotation and Intuition 71

2.7.5 Different Focus in Research 74

2.8. Key Debates and Decisions 78

2.9 Learning Theory: Social Constructivism 81

2.9.1 Assumptions of Social Constructivism 83

2.9.2 Zone of Proximal Development (ZPD) 85

2.9.2.1 Non-Knowing Growing 88

2.9.2.2 Playing and Performing 90

2.9.2.3 Imitation and Completion 93

2.9.3 Mentally of Socially-Situated 98

2.10 Writing Theory, Approach and Strategy 99

2.10.1. Expressivist/Neo-Platonic Writing Theory 100

2.10.1.1 Classroom Applications 104

2.10.2 Critical Literacies 106

2.10.3 Independent Writing 108

2.11 Conceptual Framework 111

2.12 Summary 113

CHAPTER 3 METHODOLOGY

3.1 Introduction 114

3.2 Research Paradigm 115

3.3 Research Approach and Design 117

3.4	Research Procedures	120
	3.4.1 Research Site	120
	3.4.2 Demographic Data	121
	3.4.3 Research Samples	123
	3.4.3.1 Quantitative Phase	124
	3.4.3.2 Qualitative Phase	125
	3.4.4 Research Procedures	125
3.5	Pilot Study	128
3.6	Corpus Design	130
	3.6.1 Corpus Size	132
	3.6.2 Representativeness	135
3.7	Computer Software	136
	3.7.1 LancsBox	136
	3.7.1.1 WorldLists and Keywords	138
	3.7.1.2 N-grams	140
	3.7.1.3 Concordances	141
	3.7.1.4 Log-Likelihood Calculator	143
	3.7.2 W-Matrix & UCREL Semantic Analysis System (USAS)	144
	3.7.3 Berkley Neural Parser	147
3.8	Dato Collection Method	148
	3.8.1 Document Analyses	148
	3.8.2 Semi-Structured Online Interviews	150
3.9	Data Analysis	153
	3.9.1 Comparability of the Corpora	154

3.9.2 Transcription of Corpora 156

3.9.3 Analysing Procedures 156

3.9.3.1 Obtaining Lists 158

3.9.3.2 Mistags 160

3.9.3.3 Thematic Analysis 161

3.10 Ethical Issues 165

3.10.1 Informed Consent 166

3.10.2 Anonymity 168

3.10.2.1 Parts for Anonymity 169

3.10.2.2 Replacing Names of Codes 171

3.10.2.3. Methods in Anonymity 172

3.10.3 Trustworthiness 173

3.10.3.1 Dependability 174

3.10.3.2 Credibility 175

3.10.3.2.1 Prolonged Engagement 175

3.10.3.2.2 Persistent Observation 176

3.10.3.2.3 Triangulation 176

3.10.3.2.4 Peer Debriefing 177

3.10.3.3 Transferability 178

3.10.3.4 Confirmability 179

3.11 Potential Biases and Dataset Limitations 180

3.12 Summary 181

CHAPTER 4 FINDINGS

4.1 Introduction 182

4.2	Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC)	183
4.3	Type-Token Ratio	185
4.4	Literacy Genres	188
4.5	RQ One: What are the Differences and/or Similarities in the Use of English Vocabulary in the Extended Writing Produced by Advanced Learners in Upper Primary Schools across States in Malaysia?	192
	4.5.1 Similarities and Differences in Vocabulary	192
	4.5.1.1 Functional Word (West Malaysia)	193
	4.5.1.2 Functional Word (East Malaysia)	206
	4.5.1.3 Content Words	222
	4.5.1.4 Discussion	356
	4.5.2 Similarities and Differences in Phrases (Bi-, Tri- & Qualgram)	367
	4.5.2.1 Bigram (West Malaysia)	367
	4.5.2.2 Bigram (East Malaysia)	386
	4.5.2.3 Trigram (West Malaysia)	400
	4.5.2.4 Trigram (East Malaysia)	419
	4.5.2.5 Qualgram (West Malaysia)	431
	4.5.2.6 Qualgram (East Malaysia)	449
	4.5.2.7 Discussion	459
4.6	RQ Two: What are the Differences and/or Similarities of Semantic Domains Identified form the Extended Writing Produced by Advanced Learners in Upper Primary Schools across States in Malaysia?	468
	4.6.1 Salient Semantic Domains (Malaysia)	468
	4.6.2 Similarities and Differences in the Semantic Domain (West Malaysia)	477

	4.6.3 Similarities and Differences in the Semantic Domain (East Malaysia)	516
	4.6.4 Discussion	554
4.7	RQ Three: What are the Contributing Factors on the Occurrence of Salient Vocabulary and Phrases Found in AMUPSLC?	558
	4.7.1 Academy	560
	4.7.2 Linguistic	587
	4.7.3 Social Environment Repercussions	597
	4.7.4 Insufficient Vocabulary	611
	4.7.5 Discussion	612
4.8	RQ Four: How Vocabulary and Phraseology Index Is Developed based on the AMUPSLC?	618
	4.8.1 Categorisation of Literacy Genres	618
	4.8.2 Categorisation of Salient Vocabulary and Phrases	624
	4.8.3 Discussion	646

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

5.1	Introduction	649
5.2	Potential Contributions	651
5.3	Study Implications	652
5.4	Limitations	654
5.5	Future Research Suggestions	655

REFERENCES 658

APPENDICES 733

LIST OF TABLES

Table No.		Page
1.1	Writing Objectives (Year 4 & Year 5)	11
1.2	Content and Learning Standards of Writing Skill	12
3.1	Research Questions, Data Collection Methods, Instruments and Data Analysis Technique	117
3.2	Primary School Levels and Ages upon Entry (Malaysia)	121
3.3	Demographic Data of AMUSPLC	122
3.4	Research Samples Involved	124
3.5	Semantic Domains	146
3.6	Number of Extended Writing Collected	150
3.7	Month, Time, Location and Duration of Semi-Structured Online Interviews	153
3.8	Strategies to Support Reliability in a Research	174
4.1	TTR of West and East Malaysia	187
4.2	Categorisation of Literacy Genres	189
4.3	Similarities in Functional Words: West Malaysia	194
4.4	Differences in Functional Words: West Malaysia	196
4.5	Log-Likelihood Values of West Malaysia vs. L-O-B	198
4.6	Similarities in Functional Words: East Malaysia	206
4.7	Differences in Functional Words: East Malaysia	209
4.8	Log-Likelihood Values of East Malaysia vs. L-O-B	211
4.9	Similarities in Nouns: West Malaysia	223
4.10	Differences in Nouns: West Malaysia	225
4.11	Log-Likelihood Values of West Malaysia (Nouns) vs. L-O-B	228

4.12	Similarities in Nouns: East Malaysia	243
4.13	Differences in Nouns: East Malaysia	245
4.14	Log-Likelihood Values of East Malaysia (Nouns) vs. L-O-B	247
4.15	Similarities in Verbs: West Malaysia	258
4.16	Differences in Verbs: West Malaysia	260
4.17	Log-Likelihood Values of West Malaysia (Verbs) vs. L-O-B	262
4.18	Similarities in Verbs: East Malaysia	276
4.19	Differences in Verbs: East Malaysia	278
4.20	Log-Likelihood Values of East Malaysia (Verbs) vs. L-O-B	281
4.21	Similarities in Adjectives: West Malaysia	290
4.22	Differences in Adjectives: West Malaysia	293
4.23	Log-Likelihood Values of West Malaysia (Adjectives) vs. L-O-B	295
4.24	Similarities in Adjectives: East Malaysia	310
4.25	Differences in Adjectives: East Malaysia	312
4.26	Log-Likelihood Values of East Malaysia (Adjectives) vs. L-O-B	314
4.27	Similarities in Adverbs: West Malaysia	324
4.28	Differences in Adverbs: West Malaysia	326
4.29	Log-Likelihood Values of West Malaysia (Adverbs) vs. L-O-B	328
4.30	Similarities in Adverbs: East Malaysia	341
4.31	Differences in Adverbs: East Malaysia	343
4.32	Log-Likelihood Values of East Malaysia (Adverbs) vs. L-O-B	346
4.33	Similarities in Bigrams: West Malaysia	368
4.34	Differences in Bigrams: West Malaysia	370
4.35	Log-Likelihood Values of West Malaysia (Bigram) vs. L-O-B	372

4.36	Similarities in Bigrams: East Malaysia	387
4.37	Differences in Bigrams: East Malaysia	389
4.38	Log-Likelihood Values of East Malaysia (Bigram) vs. L-O-B	391
4.39	Similarities in Trigrams: West Malaysia	401
4.40	Differences in Trigrams: West Malaysia	403
4.41	Log-Likelihood Values of West Malaysia (Trigram) vs. L-O-B	406
4.42	Similarities in Trigrams: East Malaysia	420
4.43	Differences in Trigrams: East Malaysia	422
4.44	Log-Likelihood Values of East Malaysia (Trigram) vs. L-O-B	424
4.45	Similarities in Qualgrams: West Malaysia	432
4.46	Differences in Quakgrams: West Malaysia	434
4.47	Log-Likelihood Values of West Malaysia (Qualgram) vs. L-O-B	437
4.48	Similarities in Qualgrams: East Malaysia	449
4.49	Differences in Qualgrams: East Malaysia	451
4.50	Log-Likelihood Values of East Malaysia (Qualgram) vs. L-O-B	453
4.51	Top 20 Key Semantic Domains in AMUPSLC	469
4.52	Similarities in Semantic Domains: West Malaysia	478
4.53	Differences in Semantic Domains: West Malaysia	495
4.54	Log-Likelihood Values of West Malaysia (Semantic Domain) vs. BEO6	505
4.55	Similarities in Semantic Domains: East Malaysia	517
4.56	Differences in Semantic Domains: East Malaysia	527
4.57	Log-Likelihood Values of East Malaysia (Semantic Domain) vs. BEO6	545
4.58	Category and Theme	558
4.59	Category of Salient Functional Words based on CEFR Levels	626

4.60	Category of Salient Nouns based on CEFR Levels	628
4.61	Category of Salient Verbs based on CEFR Levels	630
4.62	Category of Salient Adjectives based on CEFR Levels	633
4.63	Category of Salient Adverbs based on CEFR Levels	635
4.64	Category of Salient Bigrams based on CEFR Levels	637
4.65	Category of Salient Trigrams based on CEFR Levels	639
4.66	Category of Salient Qualgrams based on CEFR Levels	642

LIST OF FIGURES

Figure No.		Page
2.1	Categories of Corpora Classification	53
2.2	Discrepancy Relationship between the Corpus Approaches	66
2.3	Adjusted Relationship Regarding Scope and Size of Approaches	69
2.4	Optional Conceptualising of the Relationship of Two Approaches	70
2.5	Conceptual Framework	112
3.1	Research Procedures	126
4.1	Concordances “I”: West Malaysia (34 of 2,509 lines)	204
4.2	Concordances “of”: West Malaysia (34 of 2,251 lines)	204
4.3	Concordances “my”: West Malaysia (34 of 1,640 lines)	205
4.4	Concordances “I”: East Malaysia (34 of 484 lines)	217
4.5	Concordances “my”: East Malaysia (34 of 311 lines)	218
4.6	Concordances “me”: East Malaysia (34 of 80 lines)	221
4.7	Concordances “they”: East Malaysia (34 of 212 lines)	221
4.8	Concordances “day”: West Malaysia (34 of 364 lines)	235
4.9	Concordances “friends”: West Malaysia (34 of 161 lines)	236
4.10	Concordances “fox”: West Malaysia (34 of 92 lines)	239
4.11	Concordances “games”: West Malaysia (34 of 62 lines)	239
4.12	Concordances “video”: West Malaysia (34 of 52 lines)	240
4.13	Concordances “monster”: West Malaysia (34 of 51 lines)	240
4.14	Concordances “friend”: West Malaysia (34 of 41 lines)	241
4.15	Concordances “beauty”: West Malaysia (34 of 152 lines)	241
4.16	Concordances “tiger”: West Malaysia (34 of 113 lines)	242

4.17	Concordances “mother”: East Malaysia (21 of 21 lines)	254
4.18	Concordances “sister”: East Malaysia (19 of 19 lines)	255
4.19	Concordances “day”: East Malaysia (34 of 38 lines)	255
4.20	Concordances “COVID-19”: East Malaysia (34 of 81 lines)	256
4.21	Concordances “scouts”: East Malaysia (34 of 62 lines)	256
4.22	Concordances “bus”: East Malaysia (34 of 61 lines)	257
4.23	Concordances “went”: West Malaysia (34 of 322 lines)	269
4.24	Concordances “love”: West Malaysia (34 of 102 lines)	269
4.25	Concordances “got”: West Malaysia (34 of 85 lines)	272
4.26	Concordances “go”: West Malaysia (34 of 111 lines)	273
4.27	Concordances “want”: West Malaysia (34 of 43 lines)	273
4.28	Concordances “going”: West Malaysia (34 of 38 lines)	274
4.29	Concordances “saw”: West Malaysia (34 of 68 lines)	274
4.30	Concordances “get”: West Malaysia (34 of 76 lines)	275
4.31	Concordances “makes”: West Malaysia (34 of 50 lines)	275
4.32	Concordances “went”: East Malaysia (24 of 60 lines)	286
4.33	Concordances “like”: East Malaysia (16 of 16 lines)	288
4.34	Concordances “saw”: East Malaysia (14 of 14 lines)	288
4.35	Concordances “started”: East Malaysia (34 of 49 lines)	289
4.36	Concordances “cleaned”: East Malaysia (22 of 22 lines)	289
4.37	Concordances “beautiful”: West Malaysia (34 of 298 lines)	303
4.38	Concordances “favourite”: West Malaysia (30 of 75 lines)	303
4.39	Concordances “online”: West Malaysia (30 of 66 lines)	304
4.40	Concordances “happy”: West Malaysia (34 of 55 lines)	307
4.41	Concordances “violent”: West Malaysia (34 of 50 lines)	307

4.42	Concordances “aggressive”: West Malaysia (34 of 50 lines)	308
4.43	Concordances “best”: West Malaysia (34 of 60 lines)	308
4.44	Concordances “inner”: West Malaysia (31 of 31 lines)	309
4.45	Concordances “fun”: East Malaysia (9 of 16 lines)	320
4.46	Concordances “big”: East Malaysia (10 of 10 lines)	322
4.47	Concordances “excited”: East Malaysia (8 of 8 lines)	322
4.48	Concordances “online”: East Malaysia (10 of 10 lines)	322
4.49	Concordances “happy”: East Malaysia (5 of 5 lines)	323
4.50	Concordances “so”: West Malaysia (34 of 340 lines)	336
4.51	Concordances “just”: West Malaysia (34 of 95 lines)	336
4.52	Concordances “always”: West Malaysia (34 of 112 lines)	337
4.53	Concordances “how”: West Malaysia (34 of 112 lines)	337
4.54	Concordances “back”: West Malaysia (34 of 125 lines)	338
4.55	Concordances “then”: West Malaysia (34 of 105 lines)	349
4.56	Concordances “really”: West Malaysia (34 of 41 lines)	340
4.57	Concordances “up”: East Malaysia (34 of 76 lines)	352
4.58	Concordances “just”: East Malaysia (19 of 19 lines)	354
4.59	Concordances “really”: East Malaysia (16 of 16 lines)	355
4.60	Concordances “back”: East Malaysia (22 of 22 lines)	355
4.61	Concordances “together”: East Malaysia (20 of 20 lines)	356
4.62	Concordances “of the”: West Malaysia (34 of 490 lines)	380
4.63	Concordances “went to”: West Malaysia (34 of 86 lines)	382
4.64	Concordances “we went”: West Malaysia (34 of 81 lines)	383
4.65	Concordances “the monster”: West Malaysia (34 of 48 lines)	383
4.66	Concordances “video games”: West Malaysia (34 of 45 lines)	384

4.67	Concordances “it was”: West Malaysia (34 of 115 lines)	384
4.68	Concordances “I was”: West Malaysia (34 of 73 lines)	385
4.69	Concordances “the tiger”: West Malaysia (34 of 97 lines)	385
4.70	Concordances “a person”: West Malaysia (34 of 86 lines)	386
4.71	Concordances “I was”: East Malaysia (34 of 44 lines)	397
4.72	Concordances “my family”: East Malaysia (19 of 19 lines)	398
4.73	Concordances “my mother”: East Malaysia (18 of 18 lines)	398
4.74	Concordances “the bus”: East Malaysia (34 of 45 lines)	399
4.75	Concordances “the scouts”: East Malaysia (34 of 44 lines)	399
4.76	Concordances “the campsite”: East Malaysia (25 of 25 lines)	400
4.77	Concordances “we went to”: West Malaysia (34 of 66 lines)	412
4.78	Concordances “went to the”: West Malaysia (23 of 58 lines)	413
4.79	Concordances “after that we”: West Malaysia (22 of 22 lines)	415
4.80	Concordances “violent video games”: West Malaysia (21 of 21 lines)	416
4.81	Concordances “the Covid-19 pandemic”: West Malaysia (22 of 22 lines)	416
4.82	Concordances “there are many”: West Malaysia (22 of 22 lines)	417
4.83	Concordances “it was a”: West Malaysia (25 of 25 lines)	417
4.84	Concordances “a person beautiful”: West Malaysia (32 of 32 lines)	418
4.85	Concordances “makes a person”: West Malaysia (27 of 27 lines)	418
4.86	Concordances “the Covid-19 vaccine”: West Malaysia (22 of 22 lines)	419
4.87	Concordances “Neetle and Alice”: East Malaysia (9 of 9 lines)	429
4.88	Concordances “went to the”: East Malaysia (7 of 7 lines)	430
4.89	Concordances “we went to”: East Malaysia (7 of 7 lines)	430
4.90	Concordances “on the bus”: East Malaysia (15 of 15 lines)	430
4.91	Concordances “went back to”: East Malaysia (11 of 11 lines)	431

4.92	Concordances “around the world”: East Malaysia (9 of 9 lines)	431
4.93	Concordances “my family and I”: West Malaysia (17 of 34 lines)	443
4.94	Concordances “we went to the”: West Malaysia (15 of 27 lines)	443
4.95	Concordances “after that we went”: West Malaysia (12 of 12 lines)	446
4.96	Concordances “the state of nature”: West Malaysia (9 of 9 lines)	446
4.97	Concordances “I would like to”: West Malaysia (14 of 14 lines)	446
4.98	Concordances “eyes of T.J. Eckleburg”: West Malaysia (7 of 7 lines)	447
4.99	Concordances “a beautiful person is”: West Malaysia (8 of 8 lines)	447
4.100	Concordances “makes a person beautiful”: West Malaysia (21 of 21 lines)	447
4.101	Concordances “what makes a person”: West Malaysia (19 of 19 lines)	448
4.102	Concordances “children aged 5 to”: West Malaysia (16 of 16 lines)	448
4.103	Concordances “we went to the”: East Malaysia (5 of 5 lines)	458
4.104	Concordances “my family and I”: East Malaysia (4 of 4 lines)	458
4.105	Concordances “I like to do”: East Malaysia (4 of 4 lines)	458
4.106	Concordances “to stay at home”: East Malaysia (7 of 7 lines)	458
4.107	Concordances “organised a camping trip”: East Malaysia (7 of 7 lines)	459
4.108	Concordances “got on the bus”: East Malaysia (6 of 6 lines)	459
4.109	Key Semantic Domain Cloud (Malaysia)	471
4.110	Concordances of Z8: West Malaysia (34 of 16,821 lines)	511
4.111	Concordances of F1: West Malaysia (34 of 949 lines)	511
4.112	Concordances of L2: West Malaysia (34 of 636 lines)	512
4.113	Concordances of E2+: West Malaysia (34 of 138 lines)	514
4.114	Concordances of K1: West Malaysia (34 of 239 lines)	515
4.115	Concordances of O4.2+: West Malaysia (34 of 451 lines)	515
4.116	Concordances of S2: West Malaysia (34 of 453 lines)	516

4.117	Concordances of K1: East Malaysia (34 of 213 lines)	550
4.118	Concordances of Z8: East Malaysia (34 of 1,350 lines)	552
4.119	Concordances of S4: East Malaysia (34 of 186 lines)	553
4.120	Concordances of F1: East Malaysia (34 of 155 lines)	553
4.121	Concordances of M1: East Malaysia (34 of 271 lines)	554

LIST OF ABBREVIATIONS

AdvP	Adverb Phrase
AMUSPLC	Advanced Malaysian Upper Primary School Learners Corpus
AP	Adjective Phrase
BNC	British National Corpus
CBA	Corpus-Based Approach
CDA	Corpus-Driven Approach
CEFR	Common European Framework of Reference
CL	Corpus Linguistics
ELT	English Language Teaching
ESL	English as a Second Language
IP	Infinite Phrase
KSSR	Standard-Based Curriculum for Primary Schools
LL	Log-Likelihood
L-O-B	Lancaster-Oslo-Bergen
NP	Noun Phrase
PP	Prepositional Phrase
PPP	Participle Phrase
PU	Phraseological Unit
SJKC	Sekolah Jenis Kebangsaan Cina
SK	Sekolah Kebangsaan
TTR	Type-Token Ratio
USAS	UCREL Semantic Analysis System
V	Verb

VP	Verb Phrase
ZPD	Zone of Proximal Development

LIST OF APPENDICES

A	Approval of Human Research Ethics Committee Sultan Idris Education University
B	Consent Feedback Form (For Research Participants)
C	Study Information Sheet (For Research Participants)
D	Approval of Educational Research Application System
E	Approval of Conducting Research across States in Malaysia
F	Learners' Profiles
G	Teachers' Profiles
H	Experts' Profiles
I	Extended Writing Samples
J	Validators' Letter
K	Interview Protocol
L	Published Scopus-Index Articles and Award
M	Vocabulary and Phraseology Index

CHAPTER 1

INTRODUCTION

1.1 Introduction

This study examined the use of vocabulary and phrases in extended writing by advanced Malaysian upper primary school learners using a corpus linguistic (henceforth, CL) approach. Malay, Malaysia's national language, is frequently spoken in conjunction with other languages such as Mandarin and Tamil. Apart from encouraging bilingualism, the Malaysian government also emphasises the importance of English as a second language in the regular school curriculum (Hashim, 2020). The ability to compose essays in English is a critical component of the curriculum. This is primarily because they serve as building blocks for measuring English language skills (Council of Europe, 2020); hence, a component of the curriculum aims at enhancing Malaysian upper primary school learners' writing abilities includes instruction and practice in the composition of guided and extended writing. Additionally, good writing skills are

frequently evaluated through guided and extended writing (Graham, 2018). They are regularly tested at all levels of examination in Malaysia because they are considered a prevalent essay form in English Language Teaching (henceforth, ELT) classrooms.

Malaysian upper primary school learners, on the other hand, lack basic writing vocabulary and phrases (Yunus et al., 2020). Moreover, rather than teaching vocabulary and phrases from textbooks, English teachers are losing about what to introduce to school learners to help them write better essays (Lam, 2019). Guided essays appear to be a form of indoctrination based on what is presented in class (Sultana, 2020). Primary school learners made a concerted effort to memorise and copy information for guided essays. Thus, instead of highlighting guided essays which are products of memorisation, this study focused on the extended writing generated by advanced Malaysian primary school learners to collect ELT-related vocabulary and phrases in national primary schools. To assist ELT, an Advanced Malaysian Upper Primary School Learners Corpus (henceforth, AMUPSLC) and a vocabulary and phraseology index were developed to expose other primary school learners to more advanced vocabulary and phrases for extended essays.

In terms of the corpus-driven approach to primary school learners' language, Lu (2019) mentioned that writing skill is best to be examined through the contextualisation of a corpus pedagogical or topical and genre-based corpora. To be specific, Flynn (2019) recounted those customised corpora based on recurring themes and genres may be beneficial for examining and investigating primary school learners' writing in order to improve their language abilities. It aids researchers with determining the meaning of words and interpreting data for the research of English. Eventually, the

present study intended to observe the vocabulary and phrases used by advanced Malaysian upper primary school learners by examining their extended writing. Fundamentally, this study employed a corpus-driven study on advanced Malaysian upper primary school learners' extended writing across every state of Malaysia. The extended writing was compared to a reference corpus dubbed the Lancaster-Oslo Bergen (henceforth, L-O-B) corpus. In this chapter, the researcher introduced the significant motivational factors for this study with a thorough review of the language practice and expanded area regarding CL conducted in Malaysia. Later, this study was situated within the means of learning and writing theories as to the theoretical framework. The primary methodological contribution of CL emphasised the importance of keyword analyses, which led to contemporary paths for CL. Last but not least, an overview of the study was given, along with the key research questions and a synopsis of the following chapters were provided.

1.2 Background of the Study

Malaysia, a multi-cultural nation with ethnicities like Malays (69.8%), Chinese (22.4%), Indians (6.8%), and indigenous peoples (1.0%), is linguistically diverse, featuring Malay (official language), Mandarin, Tamil, and indigenous languages. English, encouraged by the government, plays a significant role in education due to globalisation influences. Since independence in 1957, Malaysia's education system, influenced by British ELT practices since the early nineteenth century, has evolved considerably. English is taught as a second language in schools, underpinning national integration. The Malaysian education, spanning thirteen years across five levels (preschool to

tertiary), emphasises English proficiency, a vital subject from primary education onwards (Ministry of Education Malaysia, 2013).

In primary schools, ESL teaching focuses on developing four key language skills (listening, writing, reading, speaking). The curriculum is divided into lower primary (Year 1-3), emphasising phonics and penmanship, and upper primary (Year 4-6), which concentrates on enhancing these language skills, particularly through essay writing. However, studies like Sahputri and Kurniawan (2019) and Durrant (2022) highlighted primary learners' struggles with English vocabulary and phrases, impacting their essay writing abilities.

Language use involves coherently combining words to express ideas, emotions, and opinions, a process central to language learning. Mastery of an extensive vocabulary is crucial for developing reading and writing skills. Consequently, this study focused on extended writing in Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC). Analysing their extended writing helped in compiling a useful vocabulary and phraseology index, serving as a reference for English language teaching and learning in Malaysian primary schools.

Recent research had explored vocabulary in non-native language environments, like Spanish-speaking students learning English as a Foreign Language (EFL). Studies such as Barcroft (2021) had investigated these areas, though such research was still limited. Binhomran and Altalhab (2021) highlighted the need for more studies on vocabulary usage in primary ESL learners to aid in developing effective teaching strategies and resources.

CL has become a prominent method for linguistic analysis in the 21st century, processing large text corpora for both qualitative and quantitative analysis. A corpus is a digitised text collection, specifically compiled for research purposes, allowing for complex manipulations of text (Egbert, 2020; Dash, 2018; Harrington, 2018). It offers tools for identifying linguistic trends in large text corpora, facilitating both quantitative and qualitative research. A prime example is Goyak et al. (2021), who conducted a corpus-driven study on mental verbs in English song lyrics, utilising software like LancsBox and the UCREL Semantic Analysis System (USAS) for quantitative and qualitative analyses. Their approach exemplifies how CL combines statistical methods (e.g., Chi-square, Log-likelihood) and corpus annotations for comprehensive linguistic research.

Globally, CL appears to be a popular area of research, as evidenced by the 6,981 journals published in Scopus between 2016 and 2022. To exemplify, Rezazadegan et al. (2022), Mohamed et al. (2022), Maimaiti et al. (2022), Wang (2022), Ortega-Bueno et al. (2022), Kagan and Solopova (2022), Gladkova et al. (2022), Isaeva et al. (2022), Ivanova and Medvedeva (2022), Martinovski (2022), Kumar et al. (2022), Sun (2022), Zhang (2022), Ädel (2022), Pa and Patel (2022), Bakarola and Nasriwala (2022), Senbel (2022), Chen and Chen (2022), Goyak et al. (2021), and Darmon et al. (2021). Thus, it is evident that CL is a critical component of the study, even more so as the world advances with technology. It is a popular research trend in other countries studies. Numerous scholars have researched a variety of fields, including speech summarisation, face mask recognition, neural machine translation, and language modelling.

On the contrary, between 2016 and 2022, only 18 Malaysian journals related to CL were published in Scopus. This demonstrates the scarcity of high-quality CL research. The studies were conducted by Almanna and Gu (2021), Zaini et al. (2021), Irham and Al Umami (2021), Low et al. (2020), Kasdan et al. (2020), Joharry and Turiman (2020), Nor and Zulcafli (2020), Sadjirin et al. (2020), Ismail et al. (2020), Juan et al (2020), Tiun et al. (2020), Sun et al. (2019), Jin et al. (2019), Shamsudin and Wan Md Adnan (2019), Lee et al. (2019), Arik and Akboga (2018), Sadjirin et al. (2018), and Turiman et al. (2018). One can conclude that CL is not widely practised or immersed enough in Malaysia. Translation, Malay architectural ingenuity and identity, language patterns, terminology, Covid-19 phrases, financial English, language modelling, Chinese language, vocational vocabulary, corpus development, gender, and spoken metadiscourse were studied. Hence, there is a need for more significant CL research in Malaysia.

Recent studies underline the importance of employing CL for the effective analysis of primary school learners' essays (Chomsky, 2020; Joharry, 2021). CL's systematic approach allows for a comparative analysis of linguistic features across corpora, utilising statistical tools and computer software. Despite the growing use of CL in social sciences, research on non-native young learners, especially at the primary level, remains limited. Zhang and Yang (2021) note a focus on tertiary education participants, with scant attention to primary school students.

This gap is particularly evident in the context of Malaysian education. While Arshad et al. (2002) analysed the EMAS corpus from Malaysian secondary students, there was a lack of similar research on advanced Malaysian upper primary school

learners, especially in vocabulary and phraseology development. This study aimed to address this gap by analysing a learner corpus of written texts from these learners, employing a corpus-driven approach to examine their vocabulary and phraseology usage over time. By comparing these findings with Arshad et al. (2002), this research contributed to a more comprehensive understanding of vocabulary and phraseology development among Malaysian learners at different educational levels.

The central aim of this study was to investigate the usage of vocabulary and phrases in extended writing by advanced Malaysian upper primary school learners. While the study did not delve into teaching methodologies or resources, it aimed to enhance the awareness of English educators in non-native contexts, particularly in Malaysian national primary schools. This understanding of learners' vocabulary and phrase choices in ESL environments is crucial.

Pedagogically, this study also reflected on the researcher’s experiences as an ESL teacher in a Malaysian primary school. By providing insights into the vocabulary and phrases used by these learners, the study aimed to aid English educators in designing more effective pedagogical tools. Such tools are expected to improve teaching methodologies, aiding learners in enriching their vocabulary and phraseology for essay writing. Ultimately, this could lead to broader linguistic acquisition and heightened language awareness among primary school learners across Malaysia, offering valuable insights for English teachers in their instructional strategies.

1.3 Problem Statement

In Malaysia, English is a compulsory subject in all national primary schools, taught as a second language to students from diverse vernacular backgrounds. Despite its integration into the curriculum, research indicates a concerning trend: a considerable number of Malaysian primary school learners exhibit low English proficiency, spanning across listening, speaking, reading, and writing skills. Notably, a segment of these students is effectively illiterate in English, struggling with word recognition and pronunciation. This scenario highlights a significant proficiency gap, as documented by studies from Lim et al. (2023), Hasram et al. (2021), Tsui (2020), Harsch and Seyferth (2020), and Dalim et al. (2020). At the core of this issue is the learners' difficulty with vocabulary and phrases, crucial elements for language mastery, particularly in writing. This deficiency manifests in their writing as grammatical inaccuracies and a lack of coherent expression.

Research by Harun and Abdullah (2020) underscores this problem, pointing out that errors in tense, punctuation, vocabulary, and spelling are prevalent among these learners, with limited vocabulary knowledge and spelling capabilities being major hurdles. Thus, this study zeroed in on vocabulary and phrases as key facets of extended writing, aiming to explore the underlying causes of these issues and develop targeted educational interventions. By addressing these critical areas, the study not only seeks to improve English language acquisition among Malaysian primary school learners but also contributes to the wider discourse on second language teaching and learning strategies.

The adoption of the Common European Framework of Reference (CEFR) for languages in Malaysian primary schools has ushered in notable challenges for English teachers and learners alike, marking a significant shift from the familiar terrain of the Standard Based Curriculum for Primary Schools (KSSR). This transition, initiated in 2013, represents a pivot towards a new set of reference materials for English teaching and learning, fundamentally altering the pedagogical landscape (C. Alih et al., 2021). The shift from KSSR to CEFR has necessitated an adaptation period, as the KSSR syllabus had previously established a set of learning standards and assessment methodologies that educators had become well-versed in. The introduction of CEFR grammar brings to the forefront new linguistic elements and teaching approaches, including the use of grammatical structures like "has/have got," a component not emphasised within the KSSR framework. Despite attempts to localise CEFR materials to better suit the Malaysian context, the transition poses linguistic and methodological hurdles for both teachers and students, particularly in terms of vocabulary and phraseology acquisition.

Under KSSR, a significant emphasis was placed on writing skills, whereas the CEFR promotes a communicative approach to language learning, prioritising oral proficiency and interactive communication over traditional writing exercises (Hui & Yunus, 2023). This fundamental shift challenges the conventional focus on writing, complicating efforts to maintain a balanced competency in both written and oral English among primary school learners. The current educational challenge involves reconciling these differences and ensuring that learners can navigate the demands of both communicative proficiency and written fluency. Addressing this gap is critical, necessitating the development of innovative tools such as the Advanced Malaysian

Upper Primary School Learners Corpus (AMUPSLC) and a targeted vocabulary and phraseology index to support effective language learning under the CEFR framework.

The transition to the Common European Framework of Reference (CEFR) in Malaysian English as a Second Language (ESL) classrooms has been a significant educational reform, scrutinised by C. Alih et al. (2021) and Navarrete (2023). This shift underscores a broader challenge in educational reforms: the critical role of teacher acceptance and adaptability, a theme echoed in earlier findings by Chin et al. (2019), who emphasised that teacher willingness is a pivotal factor in the success of any new implementation. Alih et al.'s research revealed a cautiously optimistic response from Malaysian English teachers towards incorporating CEFR into English Language Teaching (ELT). Their optimism, however, was tempered by conditions for success, encapsulated in phrases like "with sufficient support" and "just give us time." Such expressions underscored that teachers' confidence in successfully implementing CEFR hinged on specific prerequisites: ample time for familiarization with the new framework, substantial support from educational authorities in terms of resources and training, and a collaborative environment among teaching professionals.

Despite the identified optimism, the study by Chin et al. (2019) and Abdul Aziz et al. (2018) pointed out significant hurdles: the lack of sufficient time and inadequate provision of necessary materials were major barriers to effectively adopting CEFR in the classroom setting. These challenges highlight the importance of providing adequate resources and support to facilitate this educational transition (Gjikolli & Gashi-Berisha, 2023). In response to these needs, the current study aimed to offer tangible resources, such as the Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC)

and a comprehensive vocabulary and phraseology index. The index was designed to equip Malaysian English primary school teachers with the necessary materials to harness the potential of CEFR in ESL writing instruction effectively, thereby addressing the critical need for support and resources in the successful implementation of CEFR in Malaysian ESL classrooms.

In CEFR, different writing objectives are offered to Malaysian primary school learners in Years 4 and 5. The writing objectives are listed in Table 1.1 below.

Table 1.1

Writing Objectives (Year 4 & Year 5)

Writing Objectives (Year 4)	Writing Objectives (Year 5)
1. use cursive handwriting in written work.	1. give detailed information about themselves.
2. explain and give reasons for simple opinions.	2. ask for, give and respond to simple advice.
3. explain and give reasons for simple opinions.	3. narrate factual events and experiences of interest.
4. make and respond to simple offers and invitations.	4. describe people, places and objects using suitable statements.
5. describe basic everyday routines.	5. connect sentences into one or two coherent paragraphs using basic coordinating conjunctions and reference pronouns.
6. describe people and objects using suitable statements.	6. use capital letters, full stops, commas in lists and question marks appropriately in independent writing at discourse level.
7. connect sentences into a coherent paragraph using basic coordinating conjunctions and reference pronouns.	7. spell a range of high frequency words accurately in independent writing.
8. use capital letters, full stops, question marks and commas in lists appropriately in guided writing at discourse level.	8. produce a plan or draft of one or two paragraphs for a familiar topic and modify this appropriately in response to feedback.
9. spell most high frequency words accurately in guided writing.	
10. produce a plan or draft of one paragraph for a familiar topic and modify this appropriately in response to feedback.	

By emphasising vocabulary and phrases as essential components of this study, it is clear that some writing objectives anticipate Malaysian primary school learners to produce writing utilising appropriate vocabulary and phrases. For instance:

- describe basic everyday routines (Year 4);
- describe people and objects using suitable statements (Year 4 & Year 5);
- spell most high frequency words accurately in guided writing (Year 4);
- produce a plan or draft of one paragraph for a familiar topic and modify this appropriately in response to feedback (Year 4 & Year 5);
- give detailed information about themselves (Year 5);
- narrate factual events and experiences of interest (Year 5);
- spell a range of high frequency words accurately in independent writing (Year 5).

The CEFR has developed certain content and learning standards for writing skills to accomplish the aforementioned writing objectives. Each of the writing skill's content and learning standards is listed in Table 1.2.

Table 1.2

Content and Learning Standards of Writing Skill

Content and Learning Standards (Year 4)	Content and Learning Standards (Year 5)
4.2 Communicate basic information intelligibly for a range of purposes in print and digital media *4.2.3 Describe basic everyday routines* *4.2.4 Describe people and objects using suitable statements*	4.2 Communicate basic information intelligibly for a range of purposes in print and digital media *4.2.1 Give detailed information about themselves* *4.2.3 Narrate factual events and experiences of interest* *4.2.4 Describe people, places and objects using suitable statements*

Content and Learning Standards (Year 4)	Content and Learning Standards (Year 5)
4.3 Communicate with appropriate language form and style for a range of purposes in print and digital media *4.3.2 Spell most high frequency words accurately in guided writing* *4.3.3 Produce a plan or draft of one paragraph for a familiar topic and modify this appropriately in response to feedback*	4.3 Communicate with appropriate language form and style for a range of purposes in print and digital media *4.3.2 Spell a range of high frequency words accurately in independent writing* *4.3.3 Produce a plan or draft of one or two paragraphs for a familiar topic and modify this appropriately in response to feedback*

The educational standards for Malaysian upper primary school learners underscore the importance of acquiring a robust vocabulary and a repertoire of phrases essential for various writing tasks, including description, spelling, narration, and both guided and extended writing exercises. Harun and Abdullah (2020) have identified a significant gap in this area, noting that learners often lack the necessary vocabulary and phrases required for these writing tasks. This deficiency directly impacts their ability to meet the established content and learning standards in English writing, presenting a formidable challenge in their educational journey.

The role of phraseology in English cannot be overstated, as it permeates every aspect of written communication (Wang et al., 2023). Mastery of noun, verb, prepositional, verbal, participial, gerund, infinitive, and appositive phrases is crucial for creating natural-sounding English prose, especially for non-native speakers. The ability to select and use the appropriate combinations of these phrases can significantly enhance the readability and authenticity of a learner's writing, making it resonate more closely with native English expression (Lei et al., 2023). This skill becomes particularly important in scenarios such as narrating events, where the choice of phrases can either facilitate understanding or create barriers to communication among peers.

Moreover, the complexity of the English language, with its vast array of words each bearing a range of meanings—from distinct to subtly nuanced—further complicates the learning process. Crane & Malloy (2021) emphasised the importance of precise language use, encouraging learners to select the most suitable phrases to express their thoughts clearly and accurately. This precision not only aids in conveying general meanings but also ensures the transmission of specific information and detailed descriptions, thereby enriching the learners' ability to communicate effectively in English. In addressing these challenges, there is a clear need for targeted educational strategies and resources that support the development of vocabulary and phraseology among Malaysian upper primary school learners, facilitating their achievement of the requisite standards in English writing.

Hasram et al. (2021) had documented a concerning trend among Malaysian upper primary school learners: an underutilisation of English vocabulary and phrases, which significantly hampers their ability to construct effective essays. This deficiency in vocabulary and phrase usage directly correlates with their overall poor or low intermediate writing proficiency. This challenge is compounded by the difficulties faced by Malaysian English primary school teachers, who find themselves at a loss when it comes to instructing their students in the use of vocabulary and phrases for both guided and extended writing tasks (Lee & Yuan, 2021; Pham & Do, 2020; Qu, 2017).

In the ELT classrooms, the primary resources for vocabulary and phrase instruction are textbooks such as "Get Smart Plus 4" (Year 4) and "English Plus 1" (Year 5). However, these materials do not incorporate additional vocabulary or phrases beyond their pages, leading to a significant gap in learners' exposure to a broader range

of language elements. This lack of exposure is a critical factor contributing to the students' struggles with gaining phraseology competence, a competency crucial for the mastery of ESL, particularly in writing contexts (Meunier, 2019).

Identifying the root causes of these educational challenges—specifically, the ineffective use of vocabulary and phrases, along with a notable scarcity of innovative teaching strategies for these linguistic components—this study aimed to bridge these gaps. By collecting the Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC) from various states across Malaysia with a focus on vocabulary and phraseology, the study seeks to offer an option. The resulting compilation of a vocabulary and phraseology index, derived from thorough data analysis, stands as a crucial reference tool for Malaysian English primary school teachers. This index was designed to equip educators with precise and contextually appropriate phrases for instructing upper primary school learners in extended writing, thereby addressing the identified linguistic issues and enhancing the quality of English language education in Malaysia.

The scholarly landscape reveals a notable scarcity of research on the use of English phrases within Malaysian primary educational settings. Between 2016 and 2021, academic focus predominantly gravitated towards secondary and tertiary levels of education, both within Malaysia and internationally. This gap in the literature is evident from studies conducted by researchers such as Nasseri (2021), Rajendran and Yunus (2021), Saber et al. (2020), Mengü and Dönmez (2019), Singh (2019), Mukhtarova et al. (2019), Anwar (2018), Mohammad Almatarneh et al. (2018), Gayle and Shimaoka (2017), and Choo et al. (2017). Consequently, this lack of focused

research on primary education has led to a deficit in relevant references and conclusions about the use of English phrases at this critical stage of language development, rendering it difficult to formulate explicit assumptions or inferences within this context. Addressing this research gap become an essential effort.

Furthermore, the development and analysis of learner corpora have been identified as beneficial for enhancing language acquisition efforts, as highlighted by Fuchs (2020). Internationally, significant work has been undertaken in countries such as China, Russia, Korea, Turkey, Italy, Germany, Spain, and Sweden, focusing on languages other than English. This global research effort contrasts sharply with the situation in Malaysia, where only a limited number of learner corpus studies have been identified, with contributions from researchers like Joharry (2021), Kashiha (2021), Azmi et al. (2021), Ang et al. (2021), and others. This scarcity underscores a critical gap in Malaysia's research on learner corpora, particularly regarding advanced Malaysian upper primary school learners—a gap this study aims to fill by establishing the Advanced Malaysian Upper Primary School Learners Corpus (AMUSPLC).

Lastly, the absence of a vocabulary and phraseology index in Malaysia has been pinpointed as a significant concern. Such indexes have been established in various countries to support the advancement of English language teaching and learning, specifically focusing on phraseology across different educational stages. The creation of these indexes in languages like English, Spanish, Thai, and Russian stands in stark contrast to the situation in Malaysia, where no such empirical study has been undertaken. This absence potentially hinders the effective teaching and learning of English phrases, emphasizing the need for a localised vocabulary and phraseology

index. This study addressed this critical issue by developing a vocabulary and phraseology index based on 560 extended writing collected from advanced Malaysian upper primary school learners, aiming to enhance the quality of English education in Malaysia.

1.4 Purpose of the Study

The purpose of this study was to identify and analyse the vocabulary and phrases used in extended writing produced by advanced Malaysian upper primary school learners in order to create the AMUPSLC and, vocabulary and phraseology index using a corpus-driven study.

1.5 Objectives of the Study

The present study has four objectives, namely:

- to compare and contrast the use of English vocabulary and phrases in the extended writing produced by advanced learners in upper primary schools across states in Malaysia.
- to compare and contrast the semantic domains identified from the extended writing produced by advanced learners in upper primary schools across states in Malaysia.

- to investigate for the contributing factors on the occurrence of salient vocabulary and phrases found in AMUPSLC.
- to develop a vocabulary and phraseology index based on the AMUPSLC.

1.6 Research Questions

This study is guided by the following research questions:

- What are the differences and/or similarities in the use of English vocabulary and phrases in the extended writing produced by advanced learners in upper primary schools across states in Malaysia?
- What are the differences and/or similarities of semantic domains identified from the extended writing produced by advanced learners in upper primary schools across states in Malaysia?
- What are the contributing factors on the occurrence of salient vocabulary and phrases found in AMUPSLC?
- How vocabulary and phraseology index is developed based on the AMUPSLC?

1.7 Significance of the Study

The present study's findings aimed to contribute to the current Malaysian expertise in the area of CL research and the use of vocabulary and phrases in extended writing

through examining advanced Malaysian upper primary school learners' extended writing. The study employed corpus-driven approach as the theoretical framework of analysis. It emphasised the essential linguistic features which were vocabulary and phrases in the collected extended writing using LancsBox and USAS. By investigating vocabulary and phrases as written in the research questions above, the researcher expanded CL used to many domains with the purpose of ensuring the reliability of empirical data interpretation. In addition, to ensure the comparability and efficacy of corpus-driven analysis, the study demonstrated that the AMUPSLC was created through the collection of 560 extended writing with a range of topics produced by advanced Malaysian upper primary school learners across 13 states and a federal territory in Malaysia. Hence, this study was a corpus-driven approach to reveal a crucial observation: the bottom-up approach improved the findings' validity and reliability.

Furthermore, the results collected were helpful for several important purposes. First and foremost, it was to inform Malaysian education policymakers on the findings of the study. The findings were helpful for policymakers as they may decide to invest or promote the corpus-driven study and the phraseology index to other national primary schools which face the same or similar problem in teaching guided and extended essays. Also, the findings were based on actual data collected and analysed. Thus, the results were valid and reliable.

Secondly, the purpose was to inform ESL practitioners regarding the corpus-driven methodology. They could use the same research method to collect the learner corpus of young learners in their schools for different language skills such as listening and speaking. Likewise, they could use the vocabulary and phraseology index produced

with the findings of this study as a reference to teach phrases for guided essays. Finally, it might help to improve the guided and extended essays among primary school learners in Malaysia.

To further enhance the significance of the study, it is recommended to involve Malaysian education policymakers and CEFR English curriculum developers. Including these stakeholders would allow for a better understanding of how the study findings could be applied to improve the current Malaysian education system. Education policymakers would be able to use the findings of the study to inform decisions regarding the curriculum and teaching methods. They could use the vocabulary and phraseology index produced with the findings to design teaching materials that specifically target the vocabulary and phrases that Malaysian primary school learners struggle with. This would ultimately lead to improved learning outcomes for students.

CEFR English curriculum developers could also benefit from the findings of the study. The study's results could be used to inform the development of more effective English language teaching strategies that cater to the needs of Malaysian primary school learners. By incorporating the corpus-driven approach and, vocabulary and phraseology index produced with the findings, the CEFR English curriculum could be designed to better equip Malaysian primary school learners with the necessary language skills for successful extended writing. Involving education policymakers and CEFR English curriculum developers in the study would enhance the significance of the research findings and contribute to the overall improvement of the Malaysian education system.

1.8 Study Limitations

There were several limitations identified in this study. Each of them was elaborated on below. First and foremost, the study collected one extended writing from each advanced Malaysian upper primary school learner. A further study may be conducted in a more extended period to collect more essays with the same or various topics to analyse the linguistics used by them.

Moreover, the researcher could not develop a learner corpus of advanced Malaysian upper primary school learners from every district of the state mainly because it was prohibitively expensive, time-consuming, and labour-intensive. Hence, this study focused only on two national primary schools with high English performance suggested by the officers of District Education Office or State Education Departments from the capital of every state. Furthermore, Malaysia is currently experiencing a series of standard operating procedures to prevent COVID-19 that turns out to be a limitation in this study. It challenged the researcher to collect data from each district smoothly.

The thirdlimitation was the selection of primary school learners. This study focused on advanced Malaysian upper primary school learners. The findings might be more useful if the researcher included Malaysian upper primary school learners with three distinct levels of English proficiency: low, intermediate, and advanced. On the other hand, due to the manageability and objectives of this study, participation was limited to advanced Malaysian upper primary school learners. Other scholars and researchers could replicate this study by considering Malaysian upper primary school

learners with varying degrees of language competency or by covering three levels simultaneously.

1.9 Operational Definition

Below are the operational definitions of the terms used in this research.

1.9.1 Vocabulary

Vocabulary is a critical component of language acquisition. School learners must improve their vocabulary to communicate their thoughts. According to Prichard and Atkins (2021), vocabulary is defined as school learners' comprehension of oral and written words, including conceptual knowledge of the terms beyond their straightforward dictionary definitions. They emphasised that vocabulary acquisition is a continual process in which school learners make links to other words, study instances of related words, and eventually use the vocabulary correctly and appropriately within the context of the sentence. Furthermore, Ponomarenko et al. (2021) defined vocabulary as a language's words, including single words, sentences, or chunks that convey meaning. In this study, the term *vocabulary* refers to advanced Malaysian upper primary school learners' words and phrases utilised in extended writing.

1.9.2 Phrases

According to Sinclair (1991), phrases are at the heart of language description. The proclivity of words to come in the desired sequence has three ramifications that call conventional understandings of language into question. To begin, there is no difference between pattern and meaning. Following that, language is organised around the idiom and open choice principles. Finally, and perhaps most importantly, there is no distinction between lexis and grammar. Following that, the operational definition of a phrase in this study is a cluster of two, three, or four words that follow a grammatical pattern and are written by advanced Malaysian upper primary school learners in their extended writing.

1.9.3 Advanced Malaysian Upper Primary School Learners

The term *advanced Malaysian upper primary school learners* refers to 560 primary school learners who attend one of the 28 national primary schools located in the capital cities of Malaysia: Kangar, Kuantan, Kota Bahru, Johor Bahru, Alor Setar, Kuala Terengganu, Malacca City, Georgetown, Seremban, Kota Kinabalu, Ipoh, Shah Alam, and Kuching, and one federal territory, Kuala Lumpur. The national primary schools with high English proficiency were identified based on data compiled by officers in various districts and states. The 560 primary school learners were all in upper primary, from Year 4 to Year 6. They demonstrated a high level of English proficiency in their schools.

The assessment of English proficiency within this context employed both summative and formative methodologies to ascertain learners' mastery levels. Specifically, the performance level required for a high degree of proficiency is set between 5 to 6, with 6 representing the maximum score achievable through classroom-based summative assessments. This scale is designed to reflect a comprehensive evaluation of a leaner's ability to apply language skills in various academic tasks, including but not limited to writing, reading comprehension, and oral communication.

For formative assessments, which are integral to monitoring student progress and providing ongoing feedback, a benchmark of 80 marks or above has been established as the criterion for achieving a grade A. This threshold is indicative of an excellent grasp of English, encompassing a learner's ability to understand and use the language effectively in real-time scenarios, demonstrating a high level of competence across all linguistic domains. Such a dual approach to assessment ensures a holistic view of a learner's proficiency, combining the insights from performance-based evaluations with continuous assessment metrics to guide and tailor educational interventions.

1.9.4 Corpus-Driven Study

Generally, a corpus-driven study employs empirical corpus data from which language features arise naturally through data analysis. The availability of prominent and representative corpora and the incorporation of computational software enables the investigation of linguistic variation from various angles. The prospective corpus is used

to generate linguistic categories that have not been recognised by corpus linguistics or scholars, as the findings are intended to be exhaustive in terms of corpus evidence (Tognini-Bonelli, 2001). The linguistics categories are formed systematically from the recurrent patterns and frequency distributions when language is used in context (Tognini-Bonelli, 2001). According to Love (2020), certain corpus-driven studies emphasised the importance of frequency evidence, particularly when studying lexical bundles, while not prioritising frequency in analysing grammar patterns. Despite this contrast, it is assumed that the primary goal of a corpus-driven study is to discover novel language features inside a corpus inductively. As a result, this study defines corpus-driven research as a technique for analysing the vocabulary and phrases used in extended writing by advanced Malaysian upper primary school learners using empirical AMULSPLC data to allow for the natural emergence of language features through data analysis. The conclusions were explored in light of the current data corpus of AMUPSLC.

1.9.5 Extended Writing

Writing encompasses a multifaceted process that integrates various elements of textual composition, including letter formation, word selection, line spacing, line justification, identification, and hyphenation, to craft cohesive and coherent prose. It is an art form that requires the meticulous combination and connection of disparate elements, usually presented in a concise format.

Building on this foundational understanding, Jo (2021) articulated writing as an intricate mental activity that involves the discovery, communication, and development of ideas into clear, structured words and paragraphs. Kang (2021) further elaborated on this by defining writing as a means of non-verbal communication, through which an individual's feelings, thoughts, desires, and plans are expressed. This expression is particularly significant in the context of ESL, where writing serves as a pivotal tool for assessing and evaluating a learner's proficiency in language construction and comprehension.

The complexity of writing, classified as an advanced skill by Chuikova (2020), necessitates a comprehensive understanding of various components such as thesis statements, supporting details, review and editing processes, organisation, content, purpose, vocabulary, audience awareness, punctuation, prediction, and procedural knowledge. Despite these challenges, producing excellent writing pieces is of paramount importance for English teachers, scholars, researchers, textbook authors, and policymakers, serving as a reference for future educational and research.

In this study, extended writing was operationally defined as an essay that exceeds 80 words, written by advanced Malaysian upper primary school learners on a range of topics chosen according to their interests. These essays were not confined to any maximum word limit, encouraging them to explore their thoughts and ideas fully. The themes for these writings were broad and inclusive, covering areas such as the world of knowledge, world of stories, and world of family and friends, ensuring a diverse spectrum of exploration and expression.

Learners were allotted 60 minutes to complete their essays, with the option to either type their responses for a softcopy submission or write them on paper for a hardcopy. This flexibility accommodated different preferences and technological accessibilities, aiming to foster a comfortable and conducive writing environment for all learners. Through this approach, the study aimed to capture the depth and breadth of learners' ability to use vocabulary and phrases logically and cohesively, providing valuable insights into their linguistic and expressive capabilities.

1.9.6 Semantic Domains

Semantic domains are a relatively recent issue in quantitative computational linguistics, even though their fundamental ideas are derived from a long-standing research orientation in structural linguistics, for example, the notion of semantic fields (Lyons, 1977). They can be checked automatically using the lexical coherence quality revealed by texts in any language. They can be usefully employed to form a semantic network to define a computational lexicon. Semantic domains are used in this study to refer to the discourse fields of the most salient vocabulary and phrases identified from AMUPSLC based on McArthur (1981)'s postulated 21 major discourse fields.

1.9.7 Vocabulary and Phraseology Index

The essence of phraseological units plays a crucial role in defining the dynamics of phraseological meaning within linguistic studies. Kunin (2005) highlighted that a

phraseological unit consists of a combination of words whose meanings undergo transformation—either completely or partially—when they are brought together. This phenomenon underscores the significance of the essence, stability, and evolving definitions of words in their interaction with other linguistic elements. Such interactions serve as a criterion for identifying phraseological units, delineating their distinct role and positioning within the language structure. Recognised for their involvement in complex semantic processes, phrases that morph into phraseological units become integral to linguistic expressions.

Expanding on this foundation, the operational definition of the vocabulary phraseology index in this study encompassed a well-organised collection of salient vocabulary, including functional and content words, as well as two-word, three-word, and four-word phrases. These elements were derived from an analysis of 560 extended writing by advanced Malaysian upper primary school learners. Enhancements to the index introduced several key features aimed at enhancing its practicality and relevance to the Malaysian English language teaching framework: a cover and foreword section that outlines the index's objectives, its compilation methodology, and contributions; categorisation by literary genres to facilitate a genre-based learning and teaching approach; organisation of vocabulary and phrases according to CEFR levels from A1 to C2, ensuring a structured progression in language proficiency; and careful alignment with the CEFR to ensure the resource's applicability and effectiveness within the local educational context. Through these comprehensive features, the vocabulary phraseology index was transformed into an invaluable, versatile tool that significantly supports the nuanced teaching and learning of English in Malaysia, adhering to both international standards and local educational needs.

1.9.8 Advanced Malaysian Upper Primary School Learners Corpus (AMUPSLC)

According to McEnery (2019), a corpus collects naturally occurring language in electronic form, including accessible written or spoken materials. It can be saved in computer folders (Baker, 2018), allowing for manipulating and exploiting linguistics data using computer applications. Corpora are the plural of the corpus (Hornby et al., 2017). To illustrate, there are numerous famous and useful corpora available for academic and research reasons, including the following: a) British National Corpus (henceforth, BNC), b) Corpus of Contemporary American English, c) Birmingham Corpus, and d) L-O-B Corpus. The purpose of this study was to construct a corpus of extended writing produced by advanced Malaysian upper primary school learners, which will be referred to as the AMUPSLC. It consisted of 560 extended writing totalling 152,187 words.

1.10 Summary

Chapter 1 of this study provides an introduction to the study. The chapter begins with an overview of the importance of vocabulary and phraseology in language learning and the challenges faced by learners in acquiring these skills. It then presents the research questions and objectives of the study, which aim to develop a comprehensive vocabulary and phraseology index for advanced Malaysian upper primary school learners and investigate the frequency and distribution of the identified words and phrases in the learner corpus. The corpus used in the study is the AMUPSLC, which

comprises extended writing written by advanced Malaysian upper primary school learners. The chapter concludes with an overview of the structure of the study and a brief description of the chapters that follow.