



UNIVERSITI PUTRA MALAYSIA

**CORPUS-BASED ANALYSIS OF LEXICAL PATTERNS IN MALAYSIAN
SECONDARY SCHOOL SCIENCE AND ENGLISH FOR SCIENCE AND
TECHNOLOGY TEXTBOOKS**

SUJATHA MENON A/P N.S. MENON

FPP 2006 16



**CORPUS-BASED ANALYSIS OF LEXICAL PATTERNS IN MALAYSIAN
SECONDARY SCHOOL SCIENCE AND ENGLISH FOR SCIENCE AND
TECHNOLOGY TEXTBOOKS**

By

SUJATHA MENON A/P N. S. MENON

**Thesis Submitted to the School of Graduate Studies, University Putra Malaysia, in
Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

June 2009



DEDICATION

This thesis is dedicated to my parents for their continuous support, love and encouragement throughout my post-graduate studies. They have been and I know will always be my pillars of support.



Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctorate of Philosophy

**CORPUS-BASED ANALYSIS OF LEXICAL PATTERNS IN MALAYSIAN
SECONDARY SCHOOL SCIENCE AND ENGLISH FOR SCIENCE AND
TECHNOLOGY TEXTBOOKS**

By

SUJATHA MENON A/P N.S. MENON

June 2009

Chairman: Associate Professor Jayakaran Mukundan, PhD

Faculty: Faculty of Educational Studies

The teaching of Science in English in Malaysia has become an issue yet both teachers and material writers are operating in the dark, as they are teaching Science in English, when it is not quite known what scientific English is actually needed in schools. The first step into looking at the type of language used in schools and required of students for the study of Science in English language and of English for Science and Technology (EST) is, to create a corpus of the language used in these subjects.

As there is no existing corpus of the language used in the teaching and learning of Science subjects, nor for the English for Science and Technology subject, this study aims to develop two corpora: one for the Science subjects and one for the English for Science and Technology subject. These corpora will be based on the language used in both the upper secondary Science textbooks and the English for Science and



Technology textbooks from two textbook zones in Malaysia. These corpora will create a reference point for scientific and English for Science and Technology lexical patterns used in existing prescribed textbooks and would also aid materials writers and curriculum builders in the process of re-designing teaching materials in future.

This thesis analyses the similarities and differences between the lexico-grammatical patterns of scientific English and ‘everyday’ general English language and between scientific English and the language used in the English for Science and Technology textbooks. As this study intends to look at word relationships which involve collocations and multi-word clusters, the concordance software, WordSmith Tools version 4.0 is used for the purpose of text analysis in this study. Through analyses of wordlists, concordance lines and keywords, the lexico-grammatical patterns of both the corpora are identified.

The findings indicated that the English for Science and Technology textbooks from both zones were quite dissimilar, not only in vocabulary loading and distribution but also in content focus. This work has also uncovered the inadequacy of the English for Science and Technology textbook to cope with the language needs of the upper secondary Science students as even though there were many patterns shared between the English for Science and Technology and Science textbooks, the language used in the English for Science and Technology textbooks lacked variety.

The Science corpus had a greater variety of phrases, prepositions and phrasal verbs than those formed in the English for Science and Technology corpus. The study also found that even though some of the collocations in the Science corpus had predictable flexible combinations, many other collocations could not be easily predicted as they were arbitrarily blocked by usage, thus creating genre specific collocations. The study also found that many of the words when in collocation or compound form often acquire extended meanings which are content specific. General English language grammar rules could not be used to infer the meanings of compound nouns as many of the elements in a compound do not retain their literal meaning and in fact acquire extended meanings.

The creation of the Science and English for Science and Technology corpora together with the work done in this thesis on the lexis and phraseology of Scientific English used in prescribed textbooks, should be used as a platform for materials writers to design and develop more relevant and accurate English for Science and Technology textbooks and materials.



Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**KAJIAN DATA KORPUS BERASASKAN CORAK LEKSIS DALAM BUKU
TEKS SAINS DAN INGGERIS UNTUK SAINS DAN TEKNOLOGI**

Oleh

SUJATHA MENON A/P N.S. MENON

June 2009

Pengerusi: Profesor Madya Jayakaran Mukundan, PhD

Fakulti: Pengajian Pendidikan

Pengajaran dan pembelajaran Sains dalam bahasa Inggeris di Malaysia menghadapi pelbagai masalah terutamanya berkenaan kesukaran para pendidik dan penulis-penulis buku teks untuk memberi kefahaman Sains dalam bahasa Inggeris. Para pengajar masih kabur tentang jenis bahasa Inggeris saintifik yang diperlukan oleh pelajar-pelajar dan yang patut diberi focus di sekolah. Langkah pertama untuk mengenali jenis bahasa saintifik yang digunakan di sekolah dan diperlukan oleh pelajar-pelajar untuk memahami pembelajaran Sains dalam bahasa Inggeris dan juga pembelajaran subjek Inggeris untuk Sains dan Teknologi (EST), adalah dengan membentuk satu sistem pangkalan data korpus berasaskan bahasa yang digunakan dalam subjek-subjek Sains dan EST.



Oleh kerana tidak ada pangkalan data korpus yang sedia ada berasaskan bahasa Inggeris yang digunakan dalam pengajaran dan pembelajaran Sains dan EST di sekolah, kajian ini bertujuan untuk membentuk pangkalan-pangkalan data korpus Sains dan Inggeris untuk Sains dan Teknologi (EST) berasaskan bahasa yang digunakan dalam buku-buku teks Sains dan EST yang dicadangkan untuk pelajar-pelajar tingkatan 4 dan 5 di dua zon buku teks di Malaysia. Pangkalan-pangkalan korpus tersebut akan menjadi pangkalan rujukan lexis dan tatabahasa yang terkandung dalam bahasa saintifik dan EST yang digunakan di sekolah. Data korpus ini juga akan membantu penulis bahan pengajaran-pembelajaran dan pembentuk kurikulum dalam proses mengkaji semula bahan-bahan pengajaran-pembelajaran Sains dan EST.

Tesis ini menganalisa perbezaan dan kesamaan corak-corak 'lexico-grammatical' di antara bahasa Inggeris saintifik dan bahasa Inggeris 'harian' dan di antara bahasa Inggeris saintifik dan bahasa Inggeris yang digunakan dalam buku teks EST. Kajian ini telah menggunakan perisian konkordans, WordSmith Tools versi 4.0, untuk menganalisa teks dan mencerpap hubungan lexis yang melibatkan kolokasi dan 'multi-word clusters'. Corak 'lexico-grammatical' kedua-dua korpus dikenalpasti melalui penganalisan senarai kata, baris-baris konkordans dan kata kunci.

Hasil kajian menunjukkan bahawa kedua-dua buku teks EST mempunyai kemampuan dan pengagihan kosakata yang jauh berbeza. Kajian ini juga menunjukkan bahawa bahasa yang sedia ada dalam buku teks EST tidak memadai untuk menolong pelajar-pelajar dalam proses pembelajaran Sains. Walaupun terdapat beberapa corak

penggunaan bahasa yang sama di antara korpus Sains dan korpus EST, korpus Sains mempunyai pelbagai jenis corak lexis berbanding dengan korpus EST.

Korpus Sains mempunyai lebih jenis frasa, kata preposisi dan ‘phrasal verbs’ berbanding dengan korpus EST. Kajian ini juga mendedahkan beberapa kombinasi kolokasi yang dapat diramalkan dengan senang (flexible) berbanding dengan beberapa kombinasi kolokasi yang tidak boleh diramalkan (fixed) disebabkan oleh penggunaan yang spesifik. Kolokasi-kolokasi ini yang tidak fleksibel merupakan kolokasi khusus bagi genre Sains. Kajian ini juga mendapati bahawa perkataan-perkataan dalam bentuk kata majmuk (compound noun) biasanya akan memperoleh makna yang lebih luas (extended meaning) atau makna yang lain. Didapati bahawa peraturan tatabahasa bahasa Inggeris tidak dapat digunakan untuk meramalkan makna kata majmuk tersebut oleh kerana beberapa elemen dalam kombinasi-kombinasi kata majmuk tersebut tidak mengekalkan maknanya yang asal.

Pembangunan korpus Sains dan korpus EST bersama-sama dengan hasil kajian ini ke atas lexis dan frasa bahasa Saintifik harus digunakan oleh para penulis bahan pengajaran-pembelajaran untuk mengkaji semula dan membina bahan pengajaran-pembelajaran EST yang lebih relevan dan lebih berkesan.

ACKNOWLEDGEMENTS

First and foremost I wish to thank the Chairman of my supervisory committee, Associate Professor Dr. Jayakaran Mukundan. This work would not have been possible without the support and encouragement of Dr. Jayakaran. He has been abundantly helpful and has assisted me in numerous ways but more than the help given to me to complete this thesis, he has given me the encouragement and opportunity to challenge myself and realize my potential.

I would also like to thank my thesis committee members, Dr. Ghazali bin Hj. Mustapha and Dr. Ismi Arif bin Ismail who have guided me through all these years and for their substantial contributions. Thank you to Wan Hafizi bin Wan Umar, who painstakingly scanned all my texts and manually corrected them. I could not have asked for a better research assistant. Many thanks to Anealka Aziz Hussin who tutored me on the intricacies of the WordSmith concordance software. I owe a debt of gratitude to many people who have helped me along the way to completion of this project. Foremost, to my employers Universiti Teknologi MARA for giving me time off to complete my thesis and for their continuous support. I cannot end without thanking my family, on whose constant encouragement and love I have relied on. I would like to thank my husband, sons and my parents for their unconditional support all these years; having given up many things for me in the process of obtaining my PhD. They have cherished with me every great moment and supported me whenever I needed it.



I CERTIFY THAT AN Examination Committee has met on **19 June 2009** to conduct the final examination of **Sujatha Menon A/P N. S. Menon** on her **Doctor of Philosophy** thesis entitled **Corpus-Based Analysis of Lexical Patterns in Malaysian Secondary School Science and English for Science and Technology(EST) Textbooks** in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Arshad Abd. Samad, PhD

Associate Professor
Faculty of Educational Studies
University Putra Malaysia
(Chairman)

Malachi Edwin Vethamani, PhD

Associate Professor
Faculty of Educational Studies
University Putra Malaysia
(Internal Examiner)

Roselan Baki, PhD

Senior Lecturer
Faculty of Educational Studies
Universiti Putra Malaysia
(Internal Examiner)

Alan Maley

Professor
Department
University
United Kingdom
(External Examiner)

BUJANG KIM HUAT, PhD

Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:



This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of **Doctor of Philosophy**. The members of the Supervisory Committee were as follows:

Jayakaran A/L A.P. Mukundan, PhD

Associate Professor

Faculty of Educational Studies

Universiti Putra Malaysia

(Chairman)

Ghazali Bin Hj. Mustapha, PhD

Faculty of Educational Studies

Universiti Putra Malaysia

(Member)

Ismi Arif Bin Ismail, PhD

Faculty of Educational Studies

Universiti Putra Malaysia

(Member)

HASANAH MOHD. GHAZALI, PhD

Professor and Dean

School of Graduate Studies

Universiti Putra Malaysia

Date: 17 July 2009



DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

SUJATHA MENON A/P N. S. MENON

Date: 22 June 2009



TABLE OF CONTENTS

	Page
DEDICATION	ii
ABSTRACT	iii
ABSTRAK	vi
ACKNOWLEDGEMENTS	ix
APPROVAL	x
DECLARATION	xii
LIST OF TABLES	xxi
LIST OF FIGURES	xxiv
LIST OF APPENDICES	xxvi
 CHAPTER	
 1 INTRODUCTION	 1
Background to Study	1
The Language of Science	6
Corpus and Its Relevance to Language Teaching	7
Problem Statement	9
General Aim	11
Aims of Research	12
Research Questions	12



	Page
Limitations of Study	13
Scope of Study	15
Significance of Study	15
Operational Definitions	16
2 LITERATURE REVIEW	24
Introduction	24
English for Specific Purposes	24
The Development of ESP	24
Classification of ESP	29
EST Research	31
Lexis and Grammar: Relationships and Functions	35
Language of Science: BICS and CALP Distinction	38
Language of Science: Lexis and Grammar	42
Compound Nouns	46
Corpus Linguistics	50
English Language Corpus Work in Malaysia	55
The Need for a Pedagogic Corpus	56
Application of Corpora	61
Frequency Levels	62
Collocation	63

	Page
Colligation	74
Semantic Prosody	77
Multi-Word Units	78
The Lexical Approach	80
Conclusion	82
3 METHODOLOGY	83
Introduction	83
Research Design	83
Content Analysis	84
Discourse Analysis	85
The Use of Textbooks	88
Corpus Size	90
Population and Sampling in the Science and English for Science and Technology (EST) Corpora	91
Reference Corpus	96
BNC Corpus	97
Instrumentation: Word Smith Tools 4.0	99
Data Collection Procedures	103
Detailed Analysis of Data	103
Research Question 1	104
Research Question 2	111



	Page
Research Question 3	114
Research Question 4	118
Conclusion	119
4 ANALYSIS OF DATA	120
Introduction	120
General Statistics: ‘Preamble’ to Research Question 1	121
EST Statistics	124
Biology Statistics	126
Chemistry Statistics	126
Physics Statistics	127
General Science Statistics	128
Science Subjects Statistics	128
Summary of General Statistics on EST and Science Textbooks	132
Research Question 1	133
Statistics of Keyword Lists	136
Similar Keywords between Form 5 Textbook Zones	137
Comparison between Zone 1 and Zone 2 Textbooks:	
Coverage of EST Keywords	140
Comparison of Keyword Lists to Vocabulary Lists	142
Word Class Categories in the Science and EST Corpora	144
Summary of Findings: Research Question 1	145

	Page
Research Question 2	147
Distribution of Technical Words in the Science Subject	
Keyword Lists	148
Distribution of Technical Words: Comparison between	
Textbook Zones	150
Distribution of Technical Words: Comparison between	
Levels (Forms)	154
Comparison of Semi-Technical Vocabulary Use	
among the Science Subjects	158
Analysis of Similar Semi-Technical Keywords Used in	
all Four Science Subjects	159
Summary of Findings: Research Question 2	167
Research Question 3	168
Collocational Concepts Revisited	172
Research Question 3: Part A	
Lexico-Grammatical Relationships: Collocation of Selected	
Keywords in the Science Corpus	175
Analysis of Noun Keywords	178
Analysis of Adjective Keywords	195
Analysis of Verb Keywords	197
Overview of Lexico-Grammatical Patterns in the Science Corpus	200
Collocational Patterns	200



	Page
Verb Phrases	203
Prepositional Use	205
Phrasal/Prepositional Verbs	206
Semantic Prosodies	207
Summary of Findings: Research Question 3: Part A	208
Research Question 3: Part B	
Lexico-Grammatical Relationships: Collocation of Selected Keywords in the EST Corpus	209
Analysis of Noun Keywords	211
Analysis of Adjective Keywords	216
Analysis of Verb Keywords	218
Overview of Lexico-Grammatical Patterns in the EST Corpus and between the Science and EST Corpora	219
Prepositional Use in the EST Corpus	220
Phrasal/Prepositional Verbs in the EST Corpus	220
Comparison of Lexico-Grammatical Patterns between the Science and EST Corpora	222
Phrasal/Prepositional Verb Use: Comparison between EST and Science Corpora	224
Summary of Findings: Research Question 3: Part B	226
Research Question 4	227
Differences between the Science and EST	



	Page
Compound Nouns	228
Problem Areas in Compound Nouns	228
Summary of Findings: Research Question 4	231
Conclusion	232
5 DISCUSSION	233
Introduction	233
Similarities and Differences in Textbook Development	234
Language of Prescribed Textbooks and Implications for EST Materials Development	237
Comparison of the Science and EST Corpora and the Pedagogical Issues Raised with Implications for EST Textbook Development	248
Conclusion	252
6 CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH	253
Introduction	253
Conclusion	253
Recommendations for Future Research	256
Conclusion	258



REFERENCES	259
APPENDICES	275
BIODATA OF STUDENT	304
LIST OF PUBLICATIONS	305



LIST OF TABLES

Table	Page
2.1. Lexis-grammar: Relationships and functions	33
3.1. Differences between content analysis and discourse analysis	86
4.1. General statistics of Zone 1 EST text	298
4.2. General statistics of Zone 2 EST text	299
4.3. TTR of EST corpus	125
4.4. General statistics of form 4 Biology text	300
4.5. General statistics of form 5 Zone 1 Biology text	300
4.6. General statistics of form 5 Zone 2 Biology text	300
4.7. General statistics of form 4 Chemistry text	301
4.8. General statistics of form 5 Zone 1 Chemistry text	301
4.9. General statistics of form 5 Zone 2 Chemistry text	301
4.10. General statistics of form 4 Physics text	302
4.11. General statistics of form 5 Zone 2 Physics text	302
4.12. General statistics of form 5 Zone 1 Physics text	302
4.13. General statistics of form 4 General Science text	303
4.14. General statistics of form 5 Zone 2 General Science text	303
4.15. General statistics of form 5 Zone 1 General Science text	303
4.16. General statistics of Science corpora- by subject	129
4.17. Number of words in keyword lists	136
4.18. Coverage of similar keywords between the form 5 zone 1 and zone 2 textbooks	137



Table	Page
4.19. Similar keywords between EST and Science subjects	141
4.20. Similar words in vocabulary list and keyword list	142
4.21. Distribution of word class in the Science corpus keyword list	145
4.22. Distribution of word class in the EST corpus keyword list	145
4.23. Results of inter-coder reliability of subject keyword lists	148
4.24. Distribution of technical words in Science subjects	149
4.25. Results of inter-coder reliability of ‘combined’ subject and form keyword lists	151
4.26. Comparison of distribution of technical words between textbook zones	152
4.27. Comparison of the technical word distribution in the form 4 and form 5 zone 1 Science subjects	154
4.28. Comparison of the technical word distribution in the form 4 and form 5 zone 2 Science subjects	157
4.29. Breakdown of semi-technical words across subjects and forms	158
4.30. Similar semi-technical words used in all four subjects	159
4.31. Semantic categories of the noun, adjective and verb keywords in the Science corpus	176
4.32. Multiple noun phrase combinations	202
4.33. Preposition use in the Science corpus	205
4.34. Phrasal/Prepositional verbs	206
4.35. Semantic categories of the noun, adjective and verb keywords in the EST corpus	209
4.36. Preposition use in the EST corpus	220
4.37. Phrasal verbs in the EST corpus	221



Table	Page
4.38. Distribution of collocational phrases by phrase types: Comparison between EST and Science corpora	222
4.39. Similarities and differences between the Science and EST collocations	223
4.40. Comparison of phrasal verb use between the EST and Science corpora	224



LIST OF FIGURES

Figure	Page
2.1. Classification of ESP	30
2.2. Swales' EST Classification	31
2.3. Iceberg representation of BICS and CALP	39
2.4. BICS and CALP Distinction	40
2.5. Four-Step Approach to Corpus Work	51
4.1. Cross Section of a Wordlist Display	122
4.2. Comparison of STTR between EST Texts	124
4.3. Comparison of TTR and STTR between all Science and EST texts	130
4.4. Comparison of Keywords in the Physics Form 5 Zone 1 and Zone 2 Textbooks	138
4.5. Comparison of Keywords in the Biology Form 5 Zone 1 and Zone 2 Textbooks	138
4.6. Comparison of Keywords in the Chemistry Form 5 Zone 1 and Zone 2 Textbooks	139
4.7. Comparison of Keywords in the General Science Form 5 Zone 1 and Zone 2 Textbooks	139
4.8. Comparison of Keywords in the EST Zone 1 and Zone 2 Textbooks	140
4.9. Cross-Section of Physics and Chemistry Corpora Concordance Lines of the Word 'Fibre'	160
4.10. Cross-Section of the Biology Corpus Concordance lines of the Word 'Fibre'	161
4.11. Collocational Diagram of the Word 'Fibre'	162
4.12. Cross-Section of General Science and Biology Corpora Concordance Lines of the Word 'Negative'	162

