The form and function of fixed collocational patterns in research articles in different academic disciplines

David James Oakey

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Education

May 2008

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from this thesis may be published without proper acknowledgement.

Acknowledgements

First of all, I would like to thank Dr Alice Deignan for supervising this thesis. I could not have wished for a more supportive supervisor and mentor, and I am grateful for her insightful comments on my ideas during our discussions; for her constructive feedback on my written drafts; and for her serene tolerance of my just-in-time draft delivery model. I also thank Professor Lynne Cameron for acting as co-supervisor in the early days.

I would especially like to thank my parents, Robert Oakey and Maureen Oakey, for their love and support during my many fleeting appearances in Leeds while working on this thesis.

I am privileged to work in the English Department at the University of Birmingham and I would like to express my gratitude here for the advice and encouragement of my present and former colleagues: Dr Pernilla Danielsson, Dr Nick Groom, Bob Holland, Professor Susan Hunston, Tim Johns, Philip King, Dr Almut Koester, Dr Judith Lamie, Dr Jeannette Littlemore, Dr Oliver Mason, and Dr Alison Sealey. I also acknowledge the award of three grants during 2003-2005 from the Marketing, Conference and Personal Development Grants Fund of the English for International Students Unit in the School of Humanities at the University of Birmingham.

I have also benefited hugely from the friendship, advice, and encouragement of many people in the UK and further afield: Dr John Barlow, Dr Gertrud Buscher, Gordon Dobie, Dr Peter Howarth, Ramesh Krishnamurthy, Professor Michael McCarthy, John McKenny, Dr Anne O'Keefe, Dr Zeynep Onat-Stelma, Dr Ute Römer, Dr Juup Stelma, the late Professor John Sinclair and the late Dr Colin Toothill. I would also like to thank the many friends who have kept me in touch with the wider world over the years: Harinder Bains, Matt Banks, Graham and Kathryn Bamforth, Steve Brayford, John and Siobhan Cutts, Kaytee DeWolfe, Jez Franks, Richard Gidwaney, Steve Issitt, Colin Jones, Julian Pettifer, Richard Mills, Richard Scully, Phil Treece, Nick Walker, Pete Wasylyk, Anthony and Jayne Welford, and Silas Wood.

Finally, I would like to thank Professor Ronald Carter and Professor John Flowerdew for agreeing to examine this thesis.

Dedication

For Ruby, with all my love.

Abstract

This thesis presents a comparative empirical study of fixed collocational patterns in written academic English. The fixed collocational pattern is a continuous string of words which is found to occur frequently in language data. This study uses lexical analysis software to study an electronic corpus of academic research articles in an attempt to identify and compare the forms and discourse functions of fixed collocational patterns in different disciplines.

In comparative studies of the language of different academic disciplines there are two ways of collecting comparable amounts of corpus data, both of which are problematic. One approach is to subdivide existing normative (Sinclair 2005) corpora in order to allow comparisons to be made between different disciplines. The amount of data in each resulting subcorpus is often unequal, however, and results might be biased in favour of the subcorpus with the greatest number of texts or tokens. The other approach is to balance the number of tokens in each subcorpus by using incomplete text samples. This can mean that individual subcorpora do not completely represent all areas of the discourse, and some fixed collocational patterns which perform discourse functions relating to these areas may as a result be missed by the researcher.

This study attempts to establish what might be a comparable amount of data by investigating fixed collocational patterns in two different comparative corpora. First it identifies fixed collocational patterns in an equal number of tokens in each discipline, i.e. an *isolexical* comparison. It then identifies fixed collocational patterns in an equal number of texts in each discipline, i.e. an *isotextual* comparison. The findings indicate that the same fixed collocational patterns are frequent in both versions of the corpus, and so what is frequent isolexically is also frequent isotextually. This suggests that an isotextual corpus is more suitable for comparative studies of the discourse functions of fixed collocational patterns, since it allows their functions to be investigated across similar numbers of communicative acts rather than across similar amounts of language.

The thesis then compares these isotextual fixed collocational patterns with the results from two previous studies of an isolexical collocational pattern, the *lexical bundle* (Biber et al. 1999), one of which (Biber 2006) used data from a different academic genre, the other (Hyland 2008) data from three academic genres. There then follows a case study of the relationship between the lexical, semantic, and textual environments of the fixed collocational pattern *in the case of the* and its discourse functions. The thesis concludes by outlining areas of future research into fixed collocational patterns which have been suggested by the results of this study.

Table of contents

Acknowledgements 2
Dedication 3
Abstract 4
Table of contents 5
List of figures 8
List of tables 10
Chapter 1 Introduction 12
1.1 Overview 12
1.2 The puzzle of why contradictions are stark 13
1.3 Frequent word combinations: restrictions on form, meaning, and use 16
1.4 Functions of fixed collocational patterns in academic English 22
1.5 Initial research questions 25
1.6 The methodological approach 25
1.7 Organisation of the thesis 27
Chapter 2 Conceptual and methodological approaches to word combinations 29
2.1 Introduction 29
2.1.1 The terminology of word combinations 30
2.2 Meaning and the combination of individual word forms 31
2.2.1 Collocation in Lexicology 31
2.2.2 Collocation in "Eastern European" Phraseology 33
2.2.3 Collocation in the Firthian tradition 35
2.3 Empirical approaches to word combinations 38
2.3.1 Collocation in the Empirical Firthian tradition 38
2.3.2 Extended empirical collocations 44
2.4 The situational contexts of word combinations in use 48
2.4.1 Word combinations in social situations: creativity and cliché 48
2.4.2 Word combinations in social situations: language processing and cognition 52
2.4.3 Word combinations and English as a lingua franca 55
2.5 Summary 58
Chapter 3 Academic English discourse 59
3.1 Introduction 59
3.1.1 Discourse analysis 59
3.2 Descriptions of written academic discourse 60
3.2.1 Hallidayan approaches to Language, Register, and Genre in Systemic Functional Linguistics 61
3.2.2 Register Analysis 62
3.2.3 Genre Analysis 64
3.3 Functions of fixed collocational patterns in academic English discourse 65
3.3.1 Functional frameworks for lexical bundles in academic discourse: Biber (2006) and Hyland (2008) 66
3.3.2 Lexical bundles and the fixed collocational patterns in this study 70
3.4 Summary 72
Chapter 4 Corpus Linguistics: a methodology and theory of language 73
4.1 Introduction 73
4.2 The methodological approach 74
4.2.1 Introduction: corpus-based and corpus-driven approaches 74

4.2.2 The limitations of annotated data 75
4.2.3 Viewing context: the limitations of concordance lines 78
4.3 Limitations of existing corpora of written academic English 83
4.3.1 The Brown corpus tradition: Brown, LOB, FLOB, FROWN and BNC 84
4.3.2 The MicroConcord B corpus 90
4.3.3 The Bank of English 90
4.4 Corpus studies of word combinations in academic discourse 92
4.4.1 Studies with a pedagogical aim 93
4.4.2 Studies of disciplinary differences 97
4.4.3 Corpus background to lexical bundles 99
4.5 The advantages of a distinction between isolexical and isotextual versions of a comparative corpus 104
4.6 Detailed research questions 107
4.7 Summary 109
Chapter 5 Methodology: working with corpora 110
5.1 Introduction 110
5.2 The design and construction of the corpus in this study 110
5.2.1 The design of the corpus 110
5.2.2 The corpus construction procedure 111
5.3 Issues in determining corpus size 114
5.3.1 Issues in determining corpus size: software 115
5.3.2 Issues in determining corpus size: symbol noise 116
5.3.3 Issues in determining corpus size: fluff 118
5.3.4 Issues in determining corpus size: extra-textual noise 120
5.4 Comparing isolexical and isotextual versions of the corpus 125
5.4.1 Texts and tokens in the isolexical and isotextual versions of the corpus 125
5.4.2 Measuring the difference between isolexical and isotextual subcorpora 130
5.5 Identification procedure for fixed collocational patterns 131
5.5.1 The *WordSmith Tools* cluster function 131
5.5.2 Quantitative identification procedure for fixed collocational patterns 136
5.5.3 The statistical significance of fixed collocational patterns 137
5.5.4 Qualitative identification procedure for fixed collocational patterns 138
5.6 Summary 142
Chapter 6 Quantitative results 143
6.1 Introduction 143
6.2 Isolexical and isotextual frequency comparison: a brief example 143
6.3 Frequency distributions for fixed collocational patterns 146
6.3.1 Isolexical and isotextual frequency distributions of fixed collocational pattern types in the research article corpus 147
6.3.2 Isolexical and isotextual frequency distributions of fixed collocational pattern tokens in the research article corpus 150
6.4 Frequency rankings for fixed collocational patterns in the corpus in different disciplines 153
6.4.1 Isolexical and isotextual frequency rankings for the top 50 3-word fixed collocational patterns in different disciplines 154
6.4.2 Rank differences for the top 50 3-word fixed collocational patterns in different disciplines 157
6.4.3 Isolexical and isotextual frequency rankings for the top 50 4-word fixed collocational patterns in different disciplines 161
6.4.4 Rank differences for the top 50 4-word fixed collocational patterns in different disciplines 164
6.5 Fixed collocational patterns occurring frequently across all disciplines .. 167
6.5.1 3-word fixed collocational patterns occurring frequently across all disciplines 167
6.5.2 4-word fixed collocational patterns occurring frequently across all disciplines 170

6.5.3 Summary 172
Chapter 7 Fixed collocational patterns and lexical bundles 174
7.1 Comparison of fixed collocational patterns with the lexical bundles of Biber (2006) and Hyland (2008) 174
7.1.1 Comparison of isotextual fixed collocational patterns with lexical bundles: a brief example 175
7.1.2 Comparison of isotextual fixed collocational patterns with lexical bundles: full results 177
7.2 Forms of 4-word fixed collocational patterns 186
7.3 Summary 190
Chapter 8 Exploratory case study: *in the case of* 192
8.1 Introduction 192
8.2 The most frequent collocation of *in the case of* 193
8.3 Lexical environment 196
8.4 Semantic environment 199
8.5 Syntactic position 199
8.6 Textual location 201
8.7 Pragmatic function 205
8.7.1 Description of procedure 206
8.7.2 Finding from previous work 208
8.7.3 Report of findings 209
8.7.4 Explanation of topic 210
8.8 Investigating the relationship between pragmatics, lexis, syntax, and text location 211
8.8.1 Pragmatics, lexis and text location 211
8.8.2 Pragmatics and syntactic location 212
8.8.3 Pragmatics and textual location 214
8.9 Summary 216
Chapter 9 Discussion 217
9.1 Introduction 217
9.2 Summary of research findings 217
9.2.1 Research question 1: isolexical and isotextual fixed collocational patterns 217
9.2.2 Research question 2: isolexical and isotextual fixed collocational patterns across disciplines 220
9.2.3 Research question 3: comparability 220
9.3 Strengths of the research 223
9.3.1 Corpus construction and methodology 223
9.3.2 Breadth of scope 223
9.4 Limitations of the research 224
9.4.1 Data and software 224
9.4.2 Statistical significance 226
9.4.3 Intuitive functional categories 226
9.5 Summary 226
Chapter 10 Implications and further research 227
10.1 Introduction 227
10.2 Corpus Linguistics 227
10.3 Academic Discourse 229
10.4 Phraseology 230
10.5 Psycholinguistics and Pedagogy 231
10.6 Conclusion 234
References 235
Corpora referred to in this study 244

Appendices 245
Appendix 1 Letter to university departments and their replies 245
Appendix 2: Summary of the isolexical and isotextual versions of the corpus 252
Appendix 3: Detailed information on the composition of the isolexical and isotextual versions of the research article corpus 253
Appendix 4: Comparison of top 50 3-word patterns: Cell Bio. 261
Appendix 5: Comparison of top 50 3-word patterns: Electrical Engineering 262
Appendix 6: Comparison of top 50 3-word patterns: Mechanical Engineering 263
Appendix 7: Comparison of top 50 3-word patterns: Nuclear Physics 264
Appendix 8: Comparison of top 50 3-word patterns: Economics 265
Appendix 9: Comparison of top 50 3-word patterns: Marketing 266
Appendix 10: Comparison of top 50 3-word patterns: Philosophy 267
Appendix 11: Comparison of top 50 3-word patterns: Sociology 268
Appendix 12: Comparison of top 50 4-word patterns: Cell Biology 269
Appendix 13: Comparison of top 50 4-word patterns: Electrical Engineering 270
Appendix 14: Comparison of top 50 4-word patterns: Mech. Eng 271
Appendix 15: Comparison of top 50 4-word patterns: Nuclear Physics 272
Appendix 16: Comparison of top 50 4-word patterns: Economics 273
Appendix 17: Comparison of top 50 4-word patterns: Marketing 274
Appendix 18:Comparison of top 50 4-word patterns: Philosophy 275
Appendix 19: Comparison of top 50 4-word patterns: Sociology 276
Appendix 20: Lexis framed by *in the case of the* 277

List of figures

Figure 1.1: Words occurring immediately to the right of *stark* in the BoE 17
Figure 1.2: Words occurring immediately to the right of *stark raving* in the BoE 20
Figure 1.3: Words occurring immediately to the right of *is in stark* in the BoE 21
Figure 1.4: Words occurring immediately to the right of *is in stark contrast* in the BoE 22
Figure 2.1: Frequent left-hand verb collocates of the noun *purse* in the Bank of English 51
Figure 3.1: Martin's framework of genre and register related to the contexts of culture and situation 62
Figure 4.1: A sentence from the BNC with and without tags 76
Figure 4.2: Output of search for XNPR (Dagneaux et al. 1998: 168) 77
Figure 4.3: Concordance of *stark* (Partington 1998: 9) in Courier New (left) and in Times New Roman (right) 80
Figure 4.4: Concordance of *evaluations* in an Economics research article, sorted 3 words to the left (Oakey 2005: 176) 82
Figure 4.5: Distribution of text samples in the Brown Corpus and the LOB Corpus. 86
Figure 4.6: comparison of three different academic writing sub-corpora in the BNC 88
Figure 4.7: Comparison of academic sub-sections of Brown, LOB, MicroConcord B, Bank of English, and the British National Corpus 92
Figure 4.8: Authors of a research article in Nuclear Physics 94
Figure 4.9: Texts and tokens in two of the subcorpora in Biber et al. (1999: 25) ... 100
Figure 4.10: Texts and tokens in the textbook genre in Biber (2006: 25) 102
Figure 4.11: Distribution of lexical bundles across functional categories in academic textbooks (Biber 2006: 164) 102
Figure 4.12: Texts and tokens in academic discourse (Hyland 2008) 103
Figure 5.1: Part of a journal article in pdf format (above right) html format (above left), and plain text format (above) 113
Figure 5.2: Folder structure for storage of the research article corpus 114

Figure 5.3: Noise from non-ASCII characters in pdf and html files 116
Figure 5.4: Extra-textual noise in plain text file caused by text from online navigation links in an html file..... 120
Figure 5.5: Extra-textual noise in plain text file caused by the header in a pdf file. 121
Figure 5.6: The isolexical version of the research article corpus compared with other academic English corpora 124
Figure 5.7: Texts and tokens in the isolexical version of the research article corpus 125
Figure 5.8: Texts and tokens in the isotextual version of the research article corpus 127
Figure 5.9: The isotextual version of the research article corpus compared with other academic English corpora 129
Figure 5.10: Regularity of noisy clusters in Philosophy articles: distribution of *the editors of the* as shown by *WordSmith Tools* plot function 134
Figure 5.11: Distribution of *in the case of the* as shown by the WordSmith Tools plot function 135
Figure 5.12: Identification of the textual location of *in the case of the* using the "source text" view in WordSmith Concord 139
Figure 5.13: Identification of the textual location of *in the case of the* using the search function of Adobe Acrobat Reader 8 139
Figure 5.14: Excel file used to facilitate the qualitative investigation of *in the case of the*..... 141
Figure 6.1: Frequencies of 3-word pattern type in the isolexical and isotextual versions of the research article corpus..... 148
Figure 6.2: Frequencies of 4-word pattern types in the isolexical and isotextual versions of the research article corpus..... 149
Figure 6.3: Frequencies of 5-word pattern types in the isolexical and isotextual versions of the research article corpus..... 149
Figure 6.4: Frequencies of 3-word pattern tokens in the isolexical and isotextual versions of the research article corpus..... 151
Figure 6.5: Frequencies of 4-word pattern tokens in the isolexical and isotextual versions of the research article corpus..... 152
Figure 6.6: Frequencies of 5-word pattern tokens in the isolexical and isotextual versions of the research article corpus..... 152
Figure 6.7: Isolexical frequency distribution of the top 50 3-word patterns 155
Figure 6.8: Isotextual frequency distribution of the top 50 3-word patterns..... 157
Figure 6.9: Isolexical frequency distribution of the top 50 4-word patterns 162
Figure 6.10: Isotextual frequency distribution of the top 50 4-word patterns..... 163
Figure 7.1: Comparison of the percentage of the top 50 4-word fixed collocational patterns in the present study with a particular form with the percentage of lexical bundles in Hyland (2008) with the same form..... 189
Figure 8.1: Words to the right of *in the case of* in the isotextual version of the corpus 193
Figure 8.2: Occurrences of *in the case of the* in isotextual subcorpora 195
Figure 8.3: Sentence position of *in the case of the* in each discipline in the isotextual version of the corpus..... 201
Figure 8.4: Textual location of *in the case of the* in each isotextual subcorpus, by percentage of text 202
Figure 8.5: Textual location of *in the case of the* in each isotextual subcorpus, by section 204
Figure 8.6: Uses of *in the case of the* in all eight disciplines 206

Figure 8.7:The relation between sentence location and use of *in the case of the* in each isotextual subcorpus 213
Figure 8.8: The relation between textual location and use of *in the case of the* in each isotextual subcorpus 215

List of tables

Table 1.1: Frequencies of single words and contiguous combinations in example 1.6 19
Table 1.2: "Impersonal verb phrases" in Jordan (1990; 1999) 23
Table 2.1: Collocational relationships of *purse* and its near-synonyms 32
Table 2.2: Comparison of Eastern European and Empirical Firthian approaches to collocation 43
Table 2.3: Examples of 4-grams from the Google Public Web Training Corpus 44
Table 2.4: Top 20 written academic formulas (Ellis & Simpson 2005) 47
Table 3.1: Comparison of the functional frameworks of Biber (2006) and Hyland (2008) 69
Table 4.1: Distribution of texts in the Journal Corpus (Hyland 1998a) 106
Table 5.1: Word counts of the same text by different software 115
Table 5.2: Effects of extra-textual noise on calculations of Sociology and Electrical Engineering article size 121
Table 5.3: Measures of the difference in size between isolexical and isotextual versions of the corpus by discipline 130
Table 5.4: Noise in frequency lists for 4-word isotextual clusters in Philosophy and Nuclear Physics (shown in italics) 133
Table 5.5: Standard macro-structure of journal article sections 142
Table 6.1: Comparison of isolexical and isotextual frequencies of 6-word fixed collocational patterns 144
Table 6.2: Frequencies of 3, 4, and 5-word pattern types and tokens occurring 100 times or more in the isolexical and isotextual versions of the corpus 146
Table 6.3: Rank differences between top 50 3-word pattern types 159
Table 6.4: Occurrences of *in this letter* and *in this paper* 160
Table 6.5: Rank differences between top 50 4-word pattern types 165
Table 6.6: Totals for 3-word pattern types occurring isolexically and isotextually in Science, non-Science, and all disciplines 167
Table 6.7: 3-word patterns occurring across all disciplines in this study, compared with the academic writing part of the BNC and the most frequent written academic formulas of Ellis and Simpson (2005) 169
Table 6.8: Totals for 4- word pattern types occurring isolexically and isotextually in Science, non-Science, and all disciplines 171
Table 6.9: 4-word patterns occurring across all disciplines in this study, compared with the academic writing part of the BNC and the most frequent written academic formulas of Ellis and Simpson (2005) 172
Table 7.1: Summary of Biber's (2006) and Hyland's (2008) functional taxonomies for lexical bundles; keys to symbols used in Tables 7.2-7.7 175
Table 7.2: Comparison of Biber's (2006) lexical bundle functions (Stance and Discourse Organizers) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study 180
Table 7.3: Comparison of Biber's (2006) lexical bundle functions (Referential 1, 2a and 2b) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study 181

Table 7.4: Comparison of Biber's (2006) lexical bundle functions (Referential 2c, 2d and 3a) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study 182
Table 7.5: Comparison of Biber's (2006) lexical bundle functions (Referential 3b, 3c, and Multifunctional reference) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study 183
Table 7.6: Comparison of Hyland's (2008) lexical bundle functions with Biber's (2006) lexical bundle functions and the fixed collocational patterns in the present study 184
Table 7.7: The "missing" fixed collocational patterns: those occurring isotextually in more than one discipline in the present study but which are not listed among Biber's or Hyland's lexical bundles 185
Table 7.8: Structural categories of lexical bundles Biber et al. 1999 and Hyland (2008) 188
Table 8.1: Head nouns of noun groups framed by *in the case of the* in the isotextual version of the corpus 198
Table 8.2: Examples of *in the case of the* used for "description of procedure" 207
Table 8.3: Examples of *in the case of the* used for "finding from previous work" ... 208
Table 8.4: Examples of *in the case of the* used for "report of findings" 209
Table 8.5: Examples of *in the case of the* used for "explanation of topic" 210
Table 8.6: The relations between pragmatics, lexis, and text location of *in the case of the* for the 10 most frequent nouns 211
Table 10.1: Fixed collocational patterns in Schmitt et al. (2004: 130) 232

Chapter 1 Introduction

1.1 Overview

This thesis presents a comparative empirical study of a subset of written academic English. It attempts to describe and compare quantitative and qualitative aspects of fixed collocational patterns observed in a corpus of research journal articles from different academic disciplines. This first chapter outlines my reasons for undertaking the study, the primary research questions it attempts to answer, and then provides a brief rationale for the methodological approach taken in order to answer these questions.

As in any empirical study of language, a decision first has to be made regarding the language feature to be studied. Once identified, occurrences of these units can be counted, classified, compared, rearranged, substituted for one another so that the researcher can say something about the language data, and possibly say something that has not been said before about the language from which the sample is taken. The basic unit of written English is the word, a commonly used unit with roles in many systems for the description of language. For example, it can variously be seen as a semantic unit with one or more meanings, a unit of grammar with a syntactic role in how sentences are put together, or a morphological unit with a form indicated by a string of letters and other characters between two spaces. Written words can, of course, be further broken down into smaller units of analysis such as morphemes, syllables, or letters, or can be combined to form larger units of analysis such as clauses, sentences, sections, and whole texts.

The language features investigated in this thesis, however, are empirical: frequently occurring strings of words which do not necessarily fit into standard semantic, grammatical or morphological descriptions. The term *fixed collocational pattern* will be used to refer to them: they appear to be fixed, as they are only identified with this form; they are collocational, since the words in the string are next to each other; they are patterns because they are found to occur frequently with this regularity of form. This study seeks to discover which fixed collocational patterns are used frequently

by writers in different academic disciplines, and it is hoped that this will add to findings from previous work on fixed collocational patterns in academic discourse.

In order to investigate these patterns, I have collected on computer a sample of research articles in the form of a corpus divided into eight subcorpora, each of which contains texts from journals from a particular academic discipline. I have used computer software to measure the frequency of occurrence of fixed collocational patterns in each subcorpus. As this study aims to investigate similarities and differences in occurrences of fixed collocational patterns in different disciplines, the subcorpora for each discipline must be comparable. A novel dual sampling methodology has therefore been employed in order to compensate for the different lengths of research articles in different disciplines. Two versions of each subcorpus are investigated, one with an equal number of *words* in each discipline, here called an 'isolexical' sample (a new coinage from "same number of words"), and the other with a equal number of *texts* in each discipline, termed here an 'isotextual' sample (from "same number of texts"). The study compares the quantitative results from both versions of the corpus with the results of three previous studies of fixed collocational patterns, all of which identified and described them in different ways. The study then concludes with a case study of where and how the most frequent isotextual collocational pattern is used in the discourse of different disciplines.

The next section will provide a rationale for the study through an extended example which arose from my work as a teacher of English for Academic Purposes (EAP).

1.2 The puzzle of why contradictions are stark

During the past decade, EAP has been an area of rapid growth in the UK higher education sector and in the English as a Foreign Language (EFL) industry worldwide. As state funding has fallen to the point where teaching UK and EU undergraduate students is often a loss-making activity, many UK universities have looked to international postgraduate students, who pay far higher tuition fees than other students, to make up the funding shortfall. The consequent 60% increase in international student numbers (including those from the EU) has been rapid, from

around 200,000 in 1995-6 to approximately 320,000 in 2004-5 (HESA 2006). These students contribute around £4 billion annually to the HE sector (*The Guardian* 2006).

Of course, the English language encountered by these students on entering academic study in the UK is often very different from the English they have encountered up to that point. Students need to engage with the language as it is used to disseminate, preserve and create knowledge in different disciplines, and subsequently use the language to participate in these activities as experts themselves. Since the 1960s, therefore, most UK universities have employed English language teachers such as myself to provide language support to international students whose first language is not English.

As part of my teaching duties I offer advice individually to students on their academic writing. This gives me the opportunity to consult with advanced users of English from across the University, and to observe at first hand the English used in a wide variety of academic disciplines. In one such consultation I noticed the following sentence in a paper by a Korean student of Law:

> Example 1.1: There is a stark contradiction here.

To me this seemed an exceptionally appropriate use of the adjective *stark* to modify the noun *contradiction*; I congratulated the student on this and asked him why he chose to describe the contradiction with this word. He said that he wanted to mean that at this point in his paper there occurred two widely different situations which could not both be true at the same time. I found, however, that it was difficult to explain to the student why to me the combined use of these two words seemed particularly apt. This is a simple sentence in which there is nothing puzzling about modifying *contradiction*, a noun, with *stark*, an adjective. This is how words belonging to the word classes "noun" and "adjective" behave according to the rules of English grammar, and this sentence is plainly grammatical.

What was more difficult to explain, however, was why I felt that his choice of *stark contradiction* was preferable to other grammatically acceptable choices, such as *clear contradiction* or *obvious contradiction*. A possible reason for this preference

may be that there are some contradictions whose meaning can only be described using the word *stark*. Or instead, there may be a special meaning of *stark* which the word only has when it means this particular type of contradiction. A third possibility is that there is an idiomatic meaning of *stark contradiction* which cannot be understood by adding the meaning of *stark* to the meaning of *contradiction,* and which cannot be replaced by *clear contradiction* or *obvious contradiction* without meaning something else.

A fourth possible reason why this combination of words appears so preferable may not relate to the semantics of the individual words or their combined use, but may be something to do with the context of situation in which *stark contradiction* is used, and the particular force of argument that such a sentence has when it is addressed to the reader of a paper on Law. Although there is a superficial similarity to the meanings of the two statements *there is a stark contradiction here* and *there are two widely different situations here which cannot both be true at the same time*, the reader's reaction to the writer's choice of one rather than the other may have important consequences for the writer. The words in Example 1.1 may be the only appropriate ones for the writer to use to say this in the context of situation, namely a Law paper; any alternative way of expressing it may be seen as inappropriate by the reader. The combination in Example 1.1 may therefore have a pragmatic function in Law texts which it does not have in texts in other academic disciplines. This may be something which Law writers know intuitively but which is rarely articulated explicitly to international students whose first language is not English.

Since my work involves looking at academic writing in many different disciplines, I was interested in to what extent frequent word combinations had similar or different meanings and functions in different academic subjects. The syntactic, semantic, and pragmatic features of the combination of *stark* and *contradiction* in Example 1.1 might be the same across different disciplines in academic English, or they may only be significant in the language of Law.

The rest of this chapter explores theory and approaches which may help explain why *stark* tends to combine with other words, and then leads to the fixed collocational patterns which will be the subject of this study.

1.3 Frequent word combinations: restrictions on form, meaning, and use

Stark is an adjective which can be used in an attributive position before a noun or as a predicate (Quirk and Greenbaum 1973: 115) as in Example 1.2 and Example 1.3 below:

> Example 1.2: His occasional use of stark lighting, often originating on the set, created a cold aspect that exposed the harsh realities of Dickens's times.

> Example 1.3: The lighting is stark: no major shadows or darkness and no style.

Stark is termed an "inherent adjective" (Quirk & Greenbaum 1973: 121) since *stark* characterises the referent of the noun, i.e. *the lighting is stark* expresses a similar meaning to *the stark lighting*. In its predicative position as in Example 1.3 it can be used with verbs such as *be, seem, appear, look, sound, taste, and feel, smell* (Swan 1995: 13; Crystal 1996: 158).

Semantically, there are several senses for *stark* which can be found in standard reference dictionaries. The *Oxford English Dictionary 2nd Edition* (1989) (OED) lists five: hard; violent, strong, rigid, and absolute. The different senses of a word derive from the contexts in which it is used, and an empirical approach to language can identify which words are used most frequently in the context of *stark* and in turn show which of these senses is used most often. The evidence from the language data in the Bank of English (see *Corpora referred to in this study* on page 244 for details of corpora referred to in this thesis), for example, shows that, among all other nouns that it could possibly premodify, *stark* has an overwhelming tendency to modify *contrast*, as in Example 1.4 and Example 1.5 below:

> Example 1.4: ... there is a stark contrast between the light and dark areas of the image.

> Example 1.5: The contrast could hardly have been more stark.

In the 450 million words of the Bank of English (BoE) there are 4,877 occurrences of *stark*. Figure 1.1 below shows that 785 of these (16%) occur immediately to the left of *contrast,* as in Example 1.4 above, while *stark* also occurs more frequently with nouns like *choice, warning*, *reality, reminder,* and *truth* than with other possible nouns.

If there is a restriction of some kind on the things that can be described as being *stark*, then it follows that there will be restrictions on the words that a user can use *stark* with. Consequently these words which can be used with *stark* will be observed to occur in corpus data with *stark* more frequently than other words. It is this kind of restriction which seems to make Example 1.1 in the previous section more likely to a native speaker than any alternate wording, since a *contradiction* is one of the things with which the use of *stark* is not restricted. This study will treat such noticeable differences in frequency of occurrence as an indicator of significance. Other approaches to calculating statistical significance of fixed collocational patterns are discussed in section 5.5.3 below.

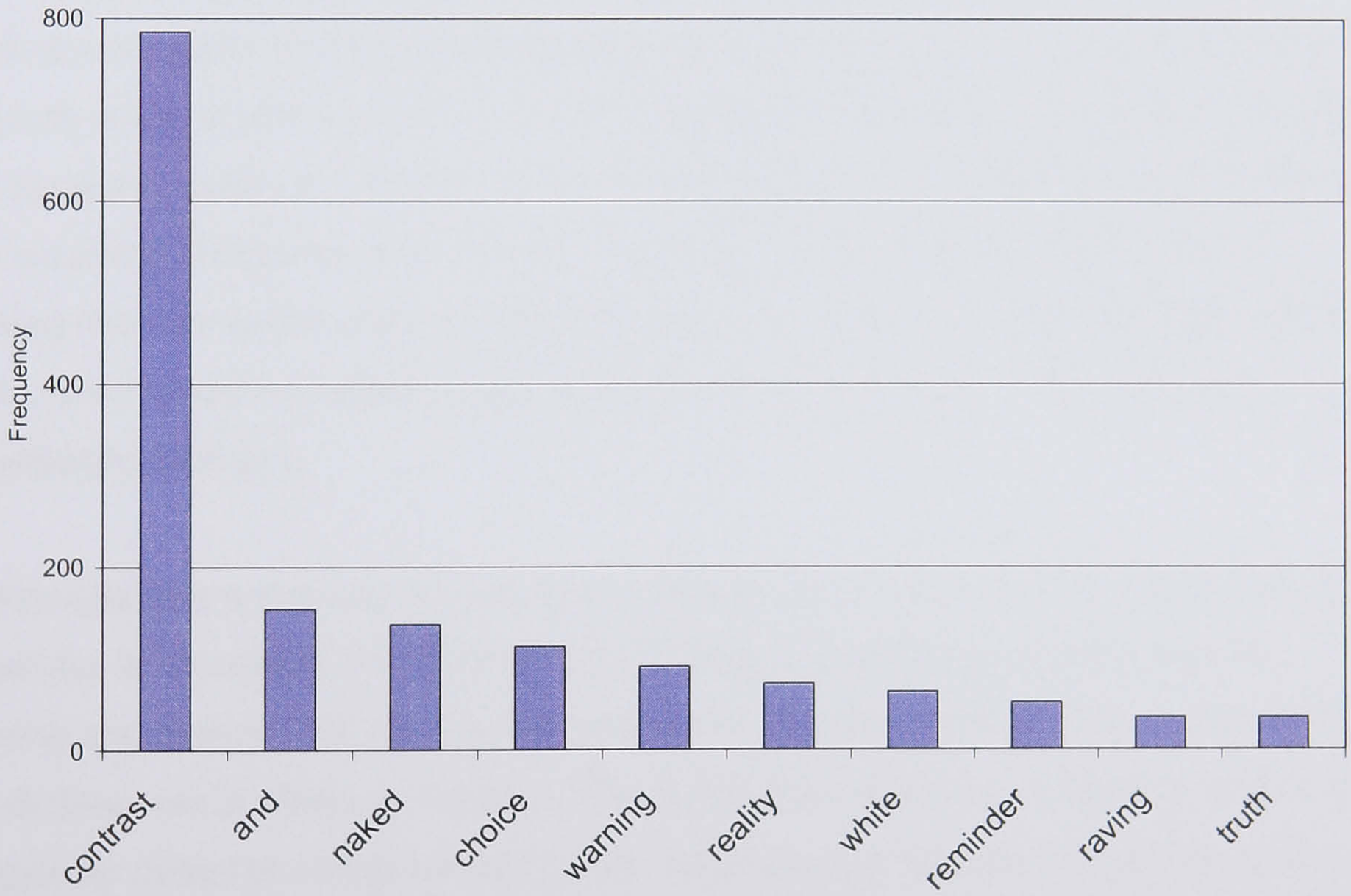


Figure 1.1: Words occurring immediately to the right of *stark* in the BoE

It can also be seen in Figure 1.1 above that in the Bank of English *stark* modifies adjectives such as *naked*, *white*, and *black* more often than others, as in Example 1.6:

> Example 1.6: Philip was so tightly wedged that when he was pulled out he left his pyjamas behind and emerged stark naked, although not seriously harmed.

Stark also seems to have the tendency, when classed as an adverb, to be restricted to modifying some adjectives more than others. The meaning of *stark*, moreover, changes according to which adjective it modifies, an observation which has implications for lexicography. A lexicographer writing a definition of *stark* would need to decide whether its tendency to occur with *naked* is so strong that *stark naked* should be treated as a single choice, or unit of meaning, and given its own separate entry in the dictionary. Both the OED and the *Collins English Dictionary 6th Edition* (2003) (CED) treat it in this way, i.e. as a headword with its own dictionary entry. The *OED* defines *stark naked* as "absolutely without clothing," while CED defines the hyphenated *stark-naked* as "completely naked".

The decision by lexicographers to give this combination its own entry in the dictionary suggests that the meaning of *stark naked* cannot be broken down into the meaning of the single word *stark* and the meaning of the single word *naked* together. The observed tendency of *stark* to occur with *naked* in corpus data thus reveals a new meaning. Someone who is *stark naked* is not wearing any fewer items of clothing than someone who is *naked.* It could thus be argued that the *stark* part of *stark naked* would be delexicalised, and have no meaning if *naked* were not immediately next to it.

A language user's decision to use *stark naked* seems not only to be motivated by semantics in the sense that in Example 1.6 *stark naked* means *Philip was not wearing any clothes,* but also by pragmatics, in that *the sight of Philip not wearing any clothes was a shock or surprise.* The writer's choice of *stark naked* rather than *completely naked* or *totally naked* has an interpersonal function in the discourse: the writer means to emphasise the shocking aspect of Philip's complete lack of clothing to the reader. A more precise definition of *stark naked* could therefore be "shockingly naked". *Stark naked* is an initial example of how single words are not always satisfactory units of meaning, and that combinations of words perform semantic and pragmatic roles in the discourse.

Example 1.6 contains 22 distinct single words and 23 different contiguous pairs of words, as shown in Table 1.1 below:

Word	Frequency		Adjacent word pair*	Frequency
he	2		although not	1
was	2		and emerged	1
although	1		behind and	1
and	1		emerged stark	1
behind	1		he left	1
emerged	1		he was	1
harmed	1		his pyjamas	1
his	1		left his	1
left	1		naked although	1
naked	1		not seriously	1
not	1		out he	1
out	1		Philip was	1
Philip	1		pulled out	1
pulled	1		pyjamas behind	1
pyjamas	1		seriously harmed	1
seriously	1		so tightly	1
so	1		stark naked	1
stark	1		that when	1
that	1		tightly wedged	1
tightly	1		was pulled	1
wedged	1		was so	1
when	1		wedged that	1
			when he	1

Table 1.1: Frequencies of single words and contiguous combinations in example 1.6
* arranged in alphabetical order

Table 1.1 was obtained by processing Example 1.6 using the *Wordsmith Tools 4* lexical analysis computer software (Scott 2005), and provides an initial indication of the methodological approach taken in this study. The table on its own may not reveal much about the properties of word pairs such as *stark naked*, but when large numbers of texts are processed using this software, frequently recurring pairs and longer combinations can be counted easily, allowing the more frequent combinations to be identified. A computer can be made to retrieve and manipulate electronically-stored corpus data and present progressively more abstract representations, i.e. from concordance lines to wordlists, and on to collocate tables and lexical frameworks (Barlow 2004b). These different perspectives are discussed in more detail in Chapter 4.

It is noticeable from corpus data that some pairs of words containing *stark* also have a tendency to combine with particular single words. For example, *stark raving* pre-modifies the adjective *mad,* as in Example 1.7:

> Example 1.7: David Aird jumped in alarm as I burst into song, rocking the boat dangerously. "For heaven's sake Bruce, have you gone stark, raving mad?"

In the Bank of English it modifies *mad* three times more frequently than its next most common word, *bonkers*, as shown in Figure 1.2 below:

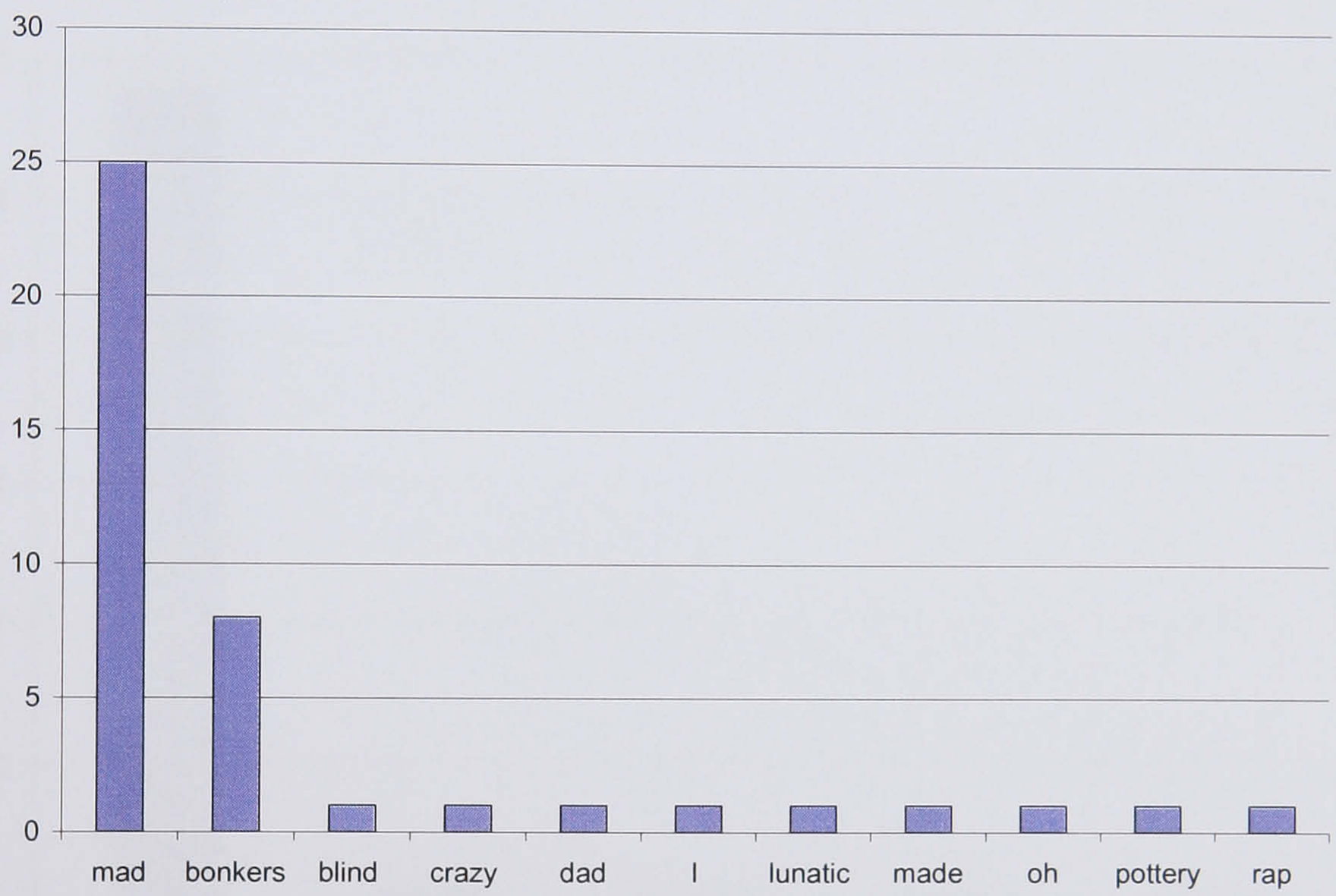


Figure 1.2: Words occurring immediately to the right of *stark raving* in the BoE

In turn, *stark raving* modifies *bonkers* eight times more frequently than it modifies other words. This means that *stark raving* is less likely to modify adjectives with similar meanings to *mad* and *bonkers*, such as *insane* or *deranged*. Such noticeable differences in the frequency of occurrence is significant. Native speaker intuition would also judge combinations such as *stark raving mad* and *stark raving bonkers* to be more marked and preferable than other possible choices such as *stark raving insane* or *stark raving deranged,* but evidence obtained by this method is less reliable than empirical evidence, as will be described in more detail in Chapter 2.

Stark also tends to be used frequently as part of a longer pattern containing a mixture of lexical and grammatical words such as that underlined in Example 1.8 below:

Example 1.8: This line of argument is in stark contrast to Drebin et al.

It can be seen from Figure 1.3 and Figure 1.4 below that in the Bank of English *is in stark* is followed by *contrast* many more times than it is followed by other words, and *is in stark contrast* in turn is followed much more often by *to* than it is by other words:

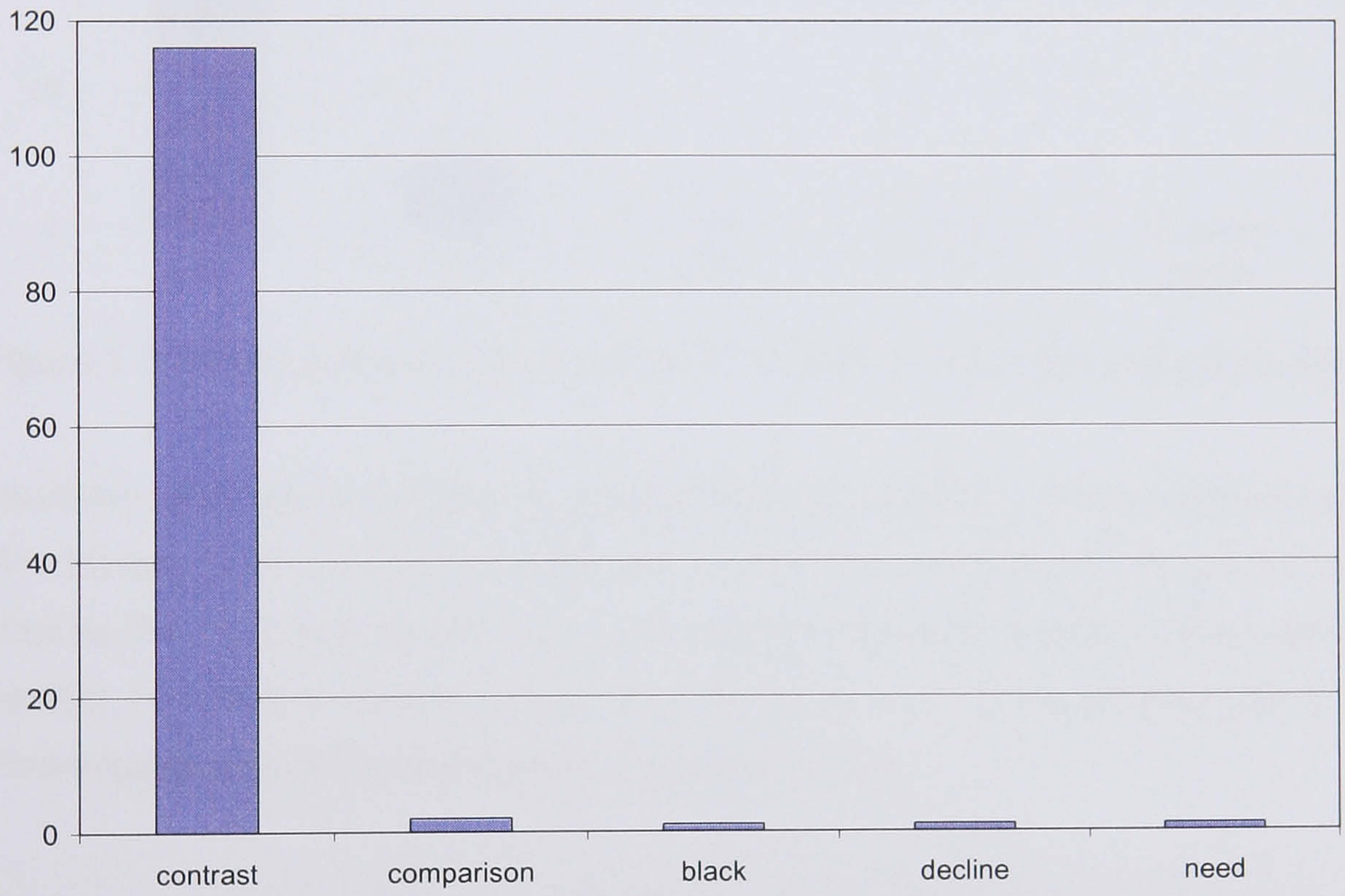


Figure 1.3: Words occurring immediately to the right of *is in stark* in the BoE

Again, the wide difference in frequency of occurrences in the corpus suggests there is a significant tendency of the words in question to be used in fixed combinations, and that combinations such as *is in stark contrast to* are likely to be fixed collocational patterns.

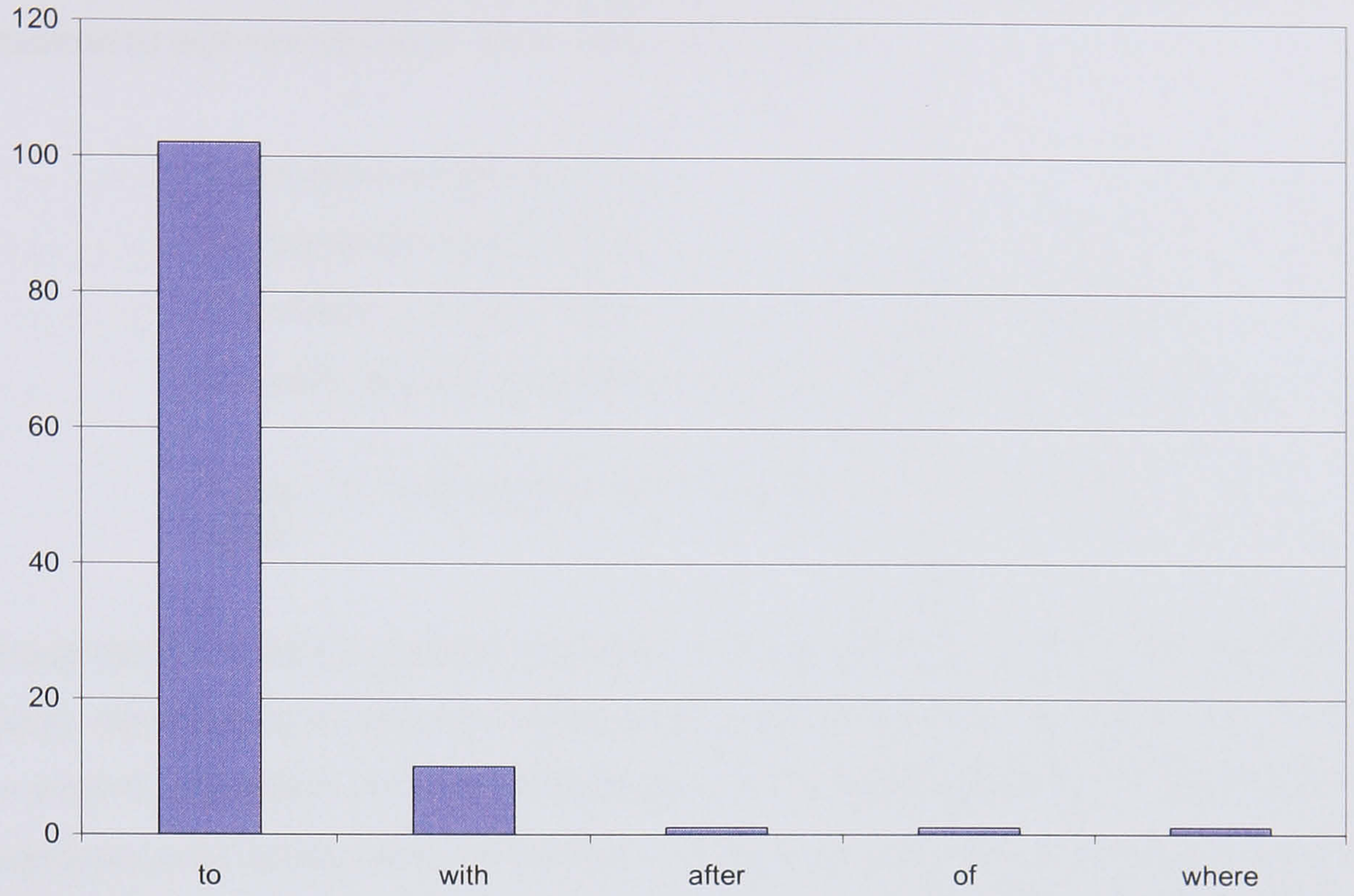


Figure 1.4: Words occurring immediately to the right of *is in stark contrast* in the BoE

Intuitively, *is in stark contrast to* seems particularly suited to written academic style. This brings up the question of pragmatics: what makes a particular sequence of word choices the most appropriate way to say something in a particular communicative context, and what is a writer doing in the text when he or she uses this pattern? These questions will be elaborated in the next section.

1.4 Functions of fixed collocational patterns in academic English

Since my interest in these fixed collocational patterns arises from my work with international students in the UK, this study accordingly focuses on comparing the form and use of such patterns in academic English. The preference by expert academic writers for some combinations over others has for many years been recognised by producers of academic writing materials, at least in the UK. Many academic writing textbooks contain lists of appropriate-sounding phrases in order to help students to write in a suitable academic style, or to structure essays in an appropriately academic way. For example, one widely-used UK academic writing

textbook (Jordan 1990: 67, 1999: 69) presents "impersonal verb phrases often associated with conclusions" as in Table 1.2 below:

It has been suggested that	
It is generally agreed that	
It is widely accepted that	(these indicate
It is now generally recognised that	stronger evidence)

Table 1.2: "Impersonal verb phrases" in Jordan (1990; 1999)

These word combinations are presented to the student as stand-alone linguistic items: introductory *it*, verb *be* in the present simple or present perfect with or without an adverb, a passive verb form followed by a clause beginning with *that*. A functional relationship for these forms in the text is specified; they have a pragmatic function, i.e. signalling the discourse structure to the reader, in this case the conclusion to a text or part of a text. They also appear to have a stylistic function: the choice of verb indicates the degree of commitment to the proposition made by the writer, and that this item, rather than a formulation such as *many people think that*, is a more appropriate academic writing style.

The *University of Leeds School of Education PhD Handbook 2007-2008*, in a similar way, matches suggested word combinations to discourse function, and advises candidates writing their theses to "show transitions carefully" (2007: 12) by using such preferred word combinations as "*the second possibility is that ...*" or "*I should now like to turn to ...*" (ibid.).

Phrases such as these provide students with apparently expert-level linguistic frames which they can immediately use to organise the discourse of their theses; such phrases might therefore be said to function as "discourse organisers". However, textbook authors who specify these phrases in their materials give little indication of why they have chosen these combinations but not others. Authors provide intuitive rather than empirical evidence for the appropriacy of these phrases for academic writing, and why they were seen as appropriate for organising certain text sections, such as conclusions. Such textbooks also do not give an indication

whether such phrases are only appropriate for use in some disciplines but not others, or whether they are appropriate in any discipline.

In previous work (Oakey 1998, 2002a) I investigated the form and function of a similar, but less fixed, collocational pattern, the *lexical phrase* proposed by Nattinger and DeCarrico (1992). I was interested in knowing more about how such units, captured in some way by materials writers and exhibited out of context, behaved in their original academic writing context. In these studies I was ultimately frustrated both by the lack of agreement on the ontological status of the lexical phrase and the unsatisfactory size and range of the academic English data available. Lexical phrases were specified for use in student essays; if items with the same form as lexical phrases were found in data from the same register of academic writing but from a different genre such as textbooks, were these forms still lexical phrases? I therefore undertook the present study of fixed collocational patterns in a corpus of my own design and construction which I intended to be an improvement on the written academic English data available, and which would give quantitative information on the frequency of fixed collocational patterns in different disciplines.

It is important to check whether results from corpora are replicable (Stubbs 2001: 124). Findings from corpora may be an artefact of that corpus, and so results need to be compared with results from different, though comparable, corpora. There are three other important studies of fixed collocational patterns against which to compare the findings from this investigation. The study by Ellis and Simpson (2005) identified "written academic formulas" using a very different methodology and data set, and will provide a measure of the replicability of my results. Two studies of "lexical bundles" (Biber 2006; Hyland 2008) provide two different functional frameworks for fixed collocational patterns in academic writing with which to compare results of my study.

The issue of register and genre, which is discussed in Chapter 3 below, is relevant to these studies. Both Hyland and Biber specified discourse functions for lexical bundles which were found in a sample of language representing the register of academic writing. Lexical bundles, however, perform discourse functions in relation to a text which belongs to a particular genre. The register data in which they identified their lexical bundle functions therefore did not contain a balanced number

of texts, and so some discourse could have been over- or under-represented in relation to the others. The data for my study, therefore, needs to be balanced for the number of words to make it comparable with these previous studies, and yet must also be balanced for the number of texts so as to make sure all possible discourse functions are represented.

1.5 Initial research questions

The first aim of the study was then to collect a corpus of complete texts from one genre, the research article, in a range of disciplines, in order to ensure that all possible fixed collocational patterns could be identified. This leads to the three initial research questions:

1. What are the most frequent fixed collocational patterns in a sample of journal articles from eight academic disciplines?

2. Do any fixed collocational patterns occur frequently across all eight academic disciplines?

3. How do these fixed collocational patterns relate to lexical bundles found by other studies?

1.6 The methodological approach

The following section briefly outlines the methodological procedure used to identify fixed collocational patterns in the data, and to ensure a balance between the number of texts and the number of tokens in each discipline.

The methodology for identifying the fixed collocational patterns in this study makes up a significant part of the original contribution of this thesis. A common approach in corpus studies is for the researcher to choose an interesting item - such as *is in stark contrast to* - and offer a description of the words that occur near it, or how it is used; such an approach reveals facts about language that have not been seen before. The

contribution of such analyses is ultimately limited, however, since the choice of *stark* above to exemplify this type of word combination was largely random: it was chosen for this chapter because of the linguistic puzzle which arose from a pedagogical encounter in the course of my work. I have no empirical indication to begin with of the significance of *is in stark contrast to* in academic English as a whole to justify investigating it in preference to other patterns.

The approach described in this thesis is instead corpus-driven (Tognini-Bonelli 2001): the patterns to be investigated were not chosen by me on the basis of intuition, or to test some other theoretical view of language already propounded, but instead presented themselves as an emergent property of the data collected for the study.

This data takes the form of a corpus of journal articles from eight different academic disciplines, partly constructed along the lines of that collected by Hyland (1998a). The corpus has several novel features. It is sizeable, at its full extent totalling a little over 40 million tokens. It is homogenous, as it contains texts from only one genre, the research article. It contains expert-level academic English, as the five journals in each section of the corpus are those judged by academics at my institution as most prestigious in their disciplines. It is a contrastive (Sinclair 2005) corpus, as its eight subcorpora are designed both to stand alone and to be compared with each other.

The most unusual aspect of the corpus in this study is that it has two versions, one *isolexical* (from "same number of words") and one *isotextual* (from "same number of texts"). The reasoning behind the distinction will be discussed in Chapter 5, but, briefly, it is difficult to compare subcorpora containing journal articles of different lengths in different disciplines. A data sample balanced for the same number of words in each discipline will contain an imbalance in the number of texts in each discipline. This imbalance in the number of texts may mean that some genre stages could be over- or under-represented in relation to other genre stages. For example, in a million-word subcorpus of 250 Electrical Engineering texts there are more introduction sections than in a million-word subcorpus of 25 Sociology texts. So, even though both subcorpora contain a million words, the Electrical Engineering subcorpus may contain more occurrences of fixed collocational patterns which are

common because they have a discourse function in introduction sections. Having two versions of the corpus should reveal more about what is a comparable amount of data.

The isolexical version of the corpus is balanced so that each subsection contains the same number of words, while the other, isotextual, version is balanced so that each subsection contains the same number of texts. The study will first compare occurrences of fixed collocational patterns between isolexical and isotextual versions to see whether the patterns that emerge are affected by the rebalancing from tokens to texts. Both versions of the corpus are potentially useful: each subsection in the isolexical version contains the same amount of *language*, and its quantitative findings are thus comparable with Biber's and Hyland's work on lexical bundles, which are defined isolexically in terms of occurrences per million words. Each subsection in the isotextual version of the corpus, on the other hand, contains the same amount of *communication*, so that the discourse functions of the patterns in the same number of texts can be compared qualitatively.

1.7 Organisation of the thesis

This thesis draws on insights from several overlapping areas of research which will be outlined in the following three chapters. Chapter 2 concerns work on Phraseology. It discusses fixed collocational patterns in relation to the field of "Eastern European" Phraseology and within the "Empirical Firthian" approach to collocation. Chapter 3 surveys earlier studies of Academic Discourse and discusses previous descriptions of the discourse functions of fixed collocational patterns. Chapter 4 describes how methodological tools from Corpus Linguistics can be used to identify fixed collocational patterns, and restates the research questions in more detail in the light of the previous three chapters. Chapter 5 then describes the application of the methodology developed in order to tackle these questions.

The next chapter presents the quantitative results of this study: Chapter 6 contains the detailed quantitative results for searches of the most frequent 3-word and 4-word fixed collocational patterns in all eight disciplines. Chapter 7 compares these patterns with those of Biber (2006), Hyland (2008), and Ellis and Simpson (2005).

Chapter 8 contains a detailed case study of the most frequently occurring fixed collocational pattern *in the case of*. Chapter 9 discusses these results and how they relate back to the original research questions. The final chapter, Chapter 10, outlines the implications of the study for the fields of Phraseology, Academic Discourse, and Corpus Linguistics, and then outlines the possible applications of the methodological approach in an agenda for further research.

Having set out the general background to the thesis and the initial research questions, we now continue by surveying the different perspectives on word combinations in order to situate the fixed collocational patterns which are the main focus of the study.

Chapter 2 Conceptual and methodological approaches to word combinations

2.1 Introduction

This chapter is an attempt to survey different approaches to the study of word combinations in order to show how the fixed collocational patterns described in this thesis relate to descriptions of word combinations in previous research. The literature on word combinations is particularly wide, and much of it has been produced as a result of research into other subjects. These include work on memory, knowledge, processing and production (Van Lancker 1974; Bolinger 1976; Pawley & Syder 1983; Wray 1999, 2002; Underwood et al. 2004), on first and second language acquisition (Hakuta 1974; Peters 1983; Wray 2002; Schmitt 2004), on dictionary design (Moon 1987; Sinclair 1987, 1991, 1996, 1998; Moon 2001, 2007), on poetry and stylistics (McIntosh 1961; Mackenzie 2000; Dillon 2006), and sociology (Kuiper 2004). Researchers in these fields are interested in different, sometimes overlapping, aspects of word combinations, and so the field is characterised by a degree of heterogeneity. Similar types of word combinations are given different names by different researchers; word combinations with very different theoretical motivations are at times referred to by the same name. While work on cognition falls outside the scope of this study, it is possible to group the literature on word combinations relevant to the present study into three areas of primary focus: meaning and the combination of individual word forms; extended empirical word combinations; and the social situations of language knowledge and use.

This chapter aims to situate the fixed collocational patterns in this study in relation to these three areas. While they emerge from the data without any prior theory, combinations of individual word forms have meaning; they are extended empirical word combinations; and as will be seen below they play a role in the social situations of language use.

2.1.1 The terminology of word combinations

The terminology associated with different theoretical approaches to word combinations reflects the many aspects of the subject. From the perspective of lexicography, which seeks to order meaning in dictionaries, meaning has over time in particular word combinations become static and their forms largely fixed; these forms are seen as having become *frozen* or *fossilized.* Cognitive approaches, namely those concerned with language processing, production and reception, instead can employ a building metaphor to hint at how such forms might be acquired, stored, and employed in language use, as in *preassembled*, *pre-formulated*, or *ready-made*, although here terms like *syntactic freezes* (Eskildsen & Cadierno 2006), in which form rather than meaning has become static, can also be found. Sociolinguistic perspectives, which highlight the role of word combinations in language use, duly focus on the repetitive, routine nature of the social situations in which they occur, as in *formulaic speech* and *conventionalized forms* (Wray 2002: 9). Computational linguists employ more literal terms, such as *multiword expressions* (Piao et al. 2003), to refer to those combinations whose ambiguous meaning cannot be reliably interpreted by natural language processing algorithms.

As already mentioned above, the abundance of terms used in the literature (Wray (2002: 9) lists over 40) arises from the different priorities of the researchers who use them. At times the combinations referred to by these terms overlap to some extent, while at others a common, shared nomenclature appears not to be possible. One potential source of confusion is that ostensibly similar items are given different names by different researchers. For example word combinations such as *the more the merrier,* and *the harder they come, the harder they fall* are for Goldberg (2003: 220) realisations of an underlying *construction,* in this case a type called a "covariational conditional", with the form *the* ____*er the* _____*er* . For Nattinger and DeCarrico however, *the* ____*er the* _____*er* is a type of *lexical phrase* called a "comparator phrasal constraint" (1992: 42).

Another possible source of confusion is that different perspectives on word combinations, such as the Firthian and "Eastern European" traditions, can contest the meaning of a single term such as *collocation,* as will be discussed in section 2.2

below. For example, in the Firthian tradition *stark naked* would be referred to as a *collocation* because of the observed high probability of *stark* and *naked* co-occurring in language use. In the "Eastern European" approach, by contrast, it would be called a *collocation* due to the fact that it meets certain syntactic, pragmatic, semantic, and lexical criteria, as it will be seen below. The Firthian idea described by the term *collocation*, moreover, would be explicitly rejected by the Eastern European tradition: "neither native speakers nor learners produce word combinations on the basis of their frequency and probability of co-occurrence" (Howarth 1998: 26).

This review of the literature therefore aims to include all work from any perspective on word combinations which is seen as relevant to the fixed collocational patterns in this study, while seeking to avoid conflating, under a particular term, distinct phenomena which are not compatible. The next section discusses approaches to word combinations which are primarily concerned with meaning and the relations between individual words.

2.2 Meaning and the combination of individual word forms

2.2.1 Collocation in Lexicology

Lexicology, the study of a store of individual words in a given language (Jackson & Zé Amvela 2000: 1), describes word combinations in terms of the semantic relations between individual words. This approach can be illustrated using the verb and noun forms of *purse*. *Purse* is an ergative verb, i.e. it has both a transitive form, in which the Subject does something, as in Example 2.1 below, and an intransitive form, as in Example 2.2 below (Hunston & Francis 2000: 183). The distinguishing feature of ergative verbs is that the object of the transitive form can also function as subject of the intransitive form.

Example 2.1: Richard pursed his lips in a long, slow whistle.

Example 2.2: William deals. He picks up his cards and his lips purse.

According to the *Collins English Dictionary* (2003: 1317) this sense of *purse* means "to contract (the mouth, lips, etc.) into a small rounded shape." By specifying *mouth* and *lips* as part of the definition of *purse*, the dictionary shows that there are

restrictions on what can be pursed in the real world. If there is a restriction on what can be pursed, then it follows that there will be restrictions on the words that an English language user can use *purse* with, and therefore these words which can be used with *purse* will be used with it more frequently than other words. The dictionary definition, however, cannot fully state the list of words which can be used with *purse*, i.e. the collocational restrictions on the use of this verb; indeed, the *New Shorter Oxford English Dictionary* (1994: 2422) in its definition of *purse* specifies a collocation with *brow* rather than with *mouth*, and refers to the final shape of something that has been pursed as "wrinkles or puckers" rather than "rounded."

In the field of lexical semantics, which investigates and describes lexical relations such as synonymy, the customary method of testing the restrictions on a word's use deriving from a particular sense is to juxtapose a set of verbs with individually similar meanings with a set of nouns with individually similar meanings. A native speaker can then be asked to judge their acceptability as a combination, similar to how a native speaker is asked to judge the grammaticality of sentences. If a native speaker decides that a particular combination is unacceptable, perhaps by saying "you could say this, but you probably wouldn't as it somehow sounds odd," then the researcher could conclude that there is a restriction on the use together of the two words. Table 2.1 below illustrates this approach by showing the acceptability (or not) of different verb-noun combinations. Combinations marked with a "+" are those which someone might intuitively judge as acceptable, those with "-" as perhaps unacceptable, and those as "?" as questionable (cf. Cruse 1986: 281):

	lips	*mouth*	*brow*	*forehead*
purse	+	?	-	-
pucker	+	+	-	-
furrow	-	-	+	?
knit	-	-	+	?

Table 2.1: Collocational relationships of *purse* and its near-synonyms

Table 2.1 suggests the existence of what Cruse (ibid.) calls "semi-systematic collocational restrictions" which act on *purse* and other near-synonyms. It could be that the meanings of *purse* and *pucker* is that they form a person's lips and mouth

into a rounded shape, but not their brow or forehead. The semantic motivation behind these restrictions might be physical: mouths and lips are able to assume rounded shapes, unlike brows and foreheads which, being wider, can only assume linear shapes. This information would be useful for a lexicographer in improving the definition of *purse,* and creating an entry for *purse* in a dictionary of collocations (cf. Benson et al. 1997).

The lexical semantic approach is not suitable for the aims of the current study since it is more concerned with collocational relations between individual word forms, rather than with strings. It is also intuitive, taking little account of observed frequency of combination, and seeks to show what is allowed or not allowed, rather than what is typical. It also often seems to be more concerned with the anomalous nature of infrequent words, such as *pucker* and *furrow*, and as such is not relevant to an empirical investigation of frequently occurring patterns such as the present study.

2.2.2 Collocation in "Eastern European" Phraseology

The next approach to word combinations originates in the linguistic traditions of the former Soviet Union and Eastern Europe, and is widely studied by Western scholars, particularly those working in lexicology and lexicography (Cowie 1998a). This field, which could be termed "Eastern European" Phraseology, also investigates the properties of an individual word which cause it to combine, or be restricted from combining, with other individual words, and categorises these combinations accordingly. The main difference between the phraseological approach and a lexical semantic approach is that these combinations are considered to be units of meaning, i.e. "word-like units which function syntactically at or below the level of the simple sentence, and sentence-like units, which function pragmatically as sayings, catchphrases, etc." (ibid. : 4). Many individual words can combine either according to syntactic and semantic rules, or in fixed expressions such as proverbs and idioms. A large number of word combinations, however, are somewhere in between: they "are much less fixed in form than idioms, but whose elements are not strictly free to associate" (ibid.). This approach is of special use to lexicographers, for whom a

fundamental question would be whether to list fixed collocations such as *stark naked* as one item or two separate words.

The methodological approach to description in this field is to intuitively select combinations of words with particular properties and categorise them according to their similarities. For example, a combination such as *stark naked* would be classed as a "restricted collocation" (Aisenstadt 1981), in that it satisfies each of four criteria, namely at the syntactic, pragmatic, semantic, and lexical levels. It is grammatically well-formed (an adverb modifying an adjective); institutionalised (somehow distinctive and memorised); partially semantically transparent, in which one element of the combination (naked) has a literal, unidiomatic meaning; and of limited commutability: *stark naked* would be acceptable to a native speaker, while combinations such as *stark nude*, *stark unclothed*, or *stark bare* would be judged as much less likely (Howarth 1996: 34-45). Similarly *stark* could replace *completely* in *completely naked*, to give *stark naked*, but *completely* in *I completely agree with you* could not be replaced by *stark* to give *I stark agree with you.**

Since they concern relations between single words, rather than the fixed strings of words investigated in the current study, most of the insights from the lexical semantic and Eastern European phraseological perspectives outlined above are less relevant here. It will be seen in section 2.3.1 below, however, that there is potential for the four-criterion framework from the Eastern European perspective to be applied to fixed collocational patterns.

One area of Phraseology that does concern strings of words is that of *phrasal verbs* and *idioms*, here used in the narrow sense to mean fully semantically opaque items like *kick the bucket* (Moon 1997: 47). Once again, however, these are not within the scope of this study. Idioms tend to be studied because their very unusualness, rather than their typicality, leads to curiosity about their origins (e.g. Funk 1950; Jack 2004). Their rarity, as O'Keeffe et al. (2007: 61) point out, often leads to idioms being presented to language learners ahead of more frequent but less salient phraseological features:

> many aspects of language are fascinating and curious in themselves, and teachers and learners know that oddity and unusualness can often be more enjoyable, learnable and memorable than the more anodyne, utilitarian elements of everyday language. (O'Keeffe et al. 2007: 61)

Idioms are commonly used in high profile areas such as sports journalism, as in example 2.3 below:

> Example 2.3: Leeds had won twice in a week and Alan Smith's four goals against Hapoel Tel Aviv had kept them in Europe. "I had thought that maybe we had turned the corner," Venables complained. "But now we are back to square one."
> (Hopps 2002)

This kind of data is easily available and much collected in electronic language corpora, with the result that work on idioms and metaphor draws on this kind of data for examples. Nonetheless, there is more evidence that idioms occur fairly infrequently, (Moon 1994, 1998) and that there is an "apparent register-constrained avoidance of idioms in expository non-fiction, in ephemera, ... in business books, and especially in academic writing" (Moon 2001: 240). Biber et al. (1999: 1025) regard them as "for the most part not common." *Back to square one* is relatively frequent, occurring once in every two million words (Moon 1996: xvii), while *turn the corner* is "regularly encountered in English," occurring between one and three times in ten million words.

The fixed collocational patterns which are the focus of this study are much more frequent and less marked than idioms, and therefore more important in descriptions of academic discourse. Of more relevance to this study is the approach to collocation based on the work of J. R. Firth, Halliday, Sinclair, and Hunston, which will be discussed next.

2.2.3 Collocation in the Firthian tradition

The area of word combinations which is most closely related to the fixed collocational patterns in this study is that of collocation. This has been an area of study in British linguistics since the work of Otto Jespersen at the beginning of the 20th century and H. E. Palmer in the 1920s (Cowie 1998b: 211). In the 1950s, J.R. Firth appropriated the anthropologist Bronislaw Malinoswski's notion of meaning as 'context of situation' :

> ... utterance and situation are bound up inextricably with each other and the context of situation is indispensable for the understanding of the words ... the utterance has no meaning except in the context of situation.
>
> (Malinowski 1923: 307)

Firth argued that words derived their meanings from their use in texts in particular situations: "the *placing* of a *text* as a constituent in a context of situation contributes to a statement of meaning since situations are set up to recognise *use*" (Firth 1957: 11, original emphasis). This means that the likelihood of whether words combine is dependent not only on individual words and the words with which they combine, but also the text in which they occur and the situation in which the text is used. This insight is of crucial importance to the present study, which looks at word combinations in a very clearly defined set of texts used in predictable situations.

While, for Firth, collocation was an abstract, linear, syntagmatic relationship between individual words which contributed to their meaning, aspects of the concept remained uncertain. Firth "was not ... the clearest of writers" (Robins 1961: 198) and his writing has been criticised as being "often allusive rather than explicit, and sometimes infuriatingly obscure on points obviously vital to the theory he was expounding" (ibid.). He was vague as to the distance collocational relationships could operate between words. It has also been noted that Firth seemed to use the term both to refer to the abstract tendency of words to co-occur, i.e. "collocation," and also to attested instances of the phenomenon observable in texts, i.e. "collocations" (some writers, e.g. Hoey (1991: 8), refer to these as "collocates"). Moreover, when referring to these observed, i.e. realised collocations, Firth is not consistent in his definition; at times he seems to be referring only to contiguous pairs of words, as in *cheese sandwich* or *strong tea*, and at other times to the relationship between words which are separated by longer stretches of text. Thus, at the time of Firth, the textual environment within which collocation relationships operated remained unclear.

Another area of uncertainty was that, while the occurrence of words together in such "habitual collocations" established a collocational level of meaning, Firth remained enigmatic about how often collocations needed to occur for them to be considered "habitual", i.e. when the frequency of occurrence of a collocation became significant

and could be recognised as in some way preferred to other possible combinations. In his famous example of "You silly ass!", "Don't be such an ass!" "What an ass he is!" the word *ass* was in "familiar and habitual company... one of the meanings of *ass* is its habitual collocation with such other words as those above quoted" (1957: 11). But while he saw collocation as the tendency of some words to occur frequently together more with some words rather than with others, he never defined what "frequent" actually meant.

Later work by Firth's student Michael Halliday, and in turn Halliday's student John Sinclair, prepared the ground for empirical studies of collocation using corpora. Halliday proposed a definition of collocation which specified both linear and statistical elements:

> the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c ... . Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more or less probable.
>
> (Halliday 1961: 276)

The key advance from Firth's notion of collocations here, as Nesselhauf (2003: 5) has pointed out, was that this defined collocation as a relationship between lexical *items* (rather than between word forms/lexemes) which were not necessarily consecutive, and which could also cross sentence boundaries. Halliday (1966: 159) reasoned that it was "not known how far collocational patterns are dependent on the structural relations into which the items enter" which formed the basis for his argument that the analysis of lexis should be performed in tandem with the analysis of grammar. This was a somewhat radical position as it implied a suspension of the basic distinction in linguistics between lexis and grammar. Halliday also pointed out that the empirical investigation of collocational relations between items "would require the study of very large samples of text" (ibid.).

The work of Firth thus prepared the way for the empirical study of word combinations using computers. This in turn led to the adoption of new units of analysis longer than a single word, such as the fixed collocational patterns investigated in the present study. This empirical approach will be discussed in section 2.3 below.

LEEDS UNIVERSITY LIBRARY

2.3 Empirical approaches to word combinations

This section discusses developments in approaches to Firthian collocation which were made possible by the development of electronic language corpora after the 1960s, and which will inform the methodological approach adopted in this study. The term used here to refer to these developments is the "Empirical Firthian" tradition. In this section the focus will be on the word combinations extracted from corpus data, rather than the issues in the design and construction of corpora, which will be discussed separately in Chapter 4, which is concerned with Corpus Linguistics.

2.3.1 Collocation in the Empirical Firthian tradition

Empirical approaches to collocation have a number of advantages over methods which rely on intuition, such as the lexical semantic and phraseological approaches described above. First, intuition reflects a user's knowledge of only those collocational relationships which they can recall having encountered in their experience of life. Some users might therefore judge some collocations acceptable whereas other users might not. Second, users recall their experiences differently. Examples of acceptable collocations elicited directly from a native speaker's intuition may only be memorable or salient to that user, and might not be recalled as memorable or salient by another user. These two limitations mean that recall data is an unreliable way of learning the most frequent acceptable collocations. Fox (1987: 146), for example, found that experimental subjects, when asked to think of collocates of a word, frequently "think more of semantic sets than of words which are actually likely to occur in the near vicinity of each other" (ibid.). Thus when asked to think of collocates of *feet*, some respondents gave other parts of the body - *legs, toes, head*, etc – while others gave *shoe, sandals, sock*; very few thought of *tall* or *high* or *long*, which are frequent collocates in corpus data. Fox also noted that once respondents were told which were the most frequent lexical collocates, they were not able to say why they had not thought of these themselves.

Instead of eliciting from native speaker intuition, empirical methods of investigating collocation use language data from a corpus, here defined as a large collection of language use, in the form of written texts or transcripts of speech, usually stored on

a computer and often designed to be a representation of the way a language is used. While in practice it is not possible to create a fully representative corpus, when data from a near-representative corpus is examined, the object of study becomes very close to being the language system itself, rather than individual users' intuitive knowledge of the language system.

The application of computing power to language study in the 1960s introduced a computer-mediated view of language as a progression of characters and spaces. This allowed researchers to measure objectively, on a large scale, both the frequency with which words occur and the frequency with which they occur with other words, and so frequently occurring collocations could be quantified.

One consequence of this development was the emergence of differing perspectives on the syntagmatic dimension of collocational relations, so that researchers adopted different views on how far apart collocational relations between words, i.e. between a node and a collocate, operate. One position is that there is no practical limit to the distance (measured in words) between the node and the collocate. *All* words in a text, i.e. "any group of sentences/utterances that functions as a unity, whether spoken or written," (Hoey 1991: 270) are in a collocational relationship with each other simply because they are all in the same text. Words combine with each other in order to create textual coherence, and these "textual collocates" (Hoey 2005: 116) make "connections between sentences across long distances of text [which] are subconsciously recognised by the reader" (1991: 219). Scott (1999) refers to words in this type of collocational relationship with another word as 'coherence collocates', similar to the collocational meaning ascribed by Sinclair (1966: 411) to words like *edition*, *bookshop*, *paper*, and *print* when they co-occur with *tome* and *paperback*.

For words which actually co-occur very close to the word *letter* – (*my*, *this*, *a*, etc.), Scott uses the term 'neighbourhood collocates'. This mirrors Halliday and Hasan's explanation of this aspect of coherence in texts: "a word that is in some way associated with another word in the preceding text, because it is a repetition of it, or is in some sense synonymous with it, or tends to occur in the same lexical environment, coheres with that word and so contributes to the texture" (Halliday & Hasan 1976: 319). This approach to collocation has been taken furthest by Hoey

(2005), who uses corpus evidence as a way of explaining how language is acquired. When a language user is repeatedly exposed to the different frequent contexts of a word, namely its collocational environments, the user's knowledge of and ability to use that word is developed, or 'primed'. This takes the Empirical Firthian approach to collocation from Corpus Linguistics to Psycholinguistics. While crucial to a full understanding of the role of word combinations in language acquisition, reception and production, the cognitive aspects of collocations is less relevant to the field of written academic discourse under investigation here. The issues relating to the investigation of the role of fixed collocational patterns in the different aspects of language cognition, however, are crucial to a broader understanding of word combinations in language, and so they will be briefly discussed in section 2.4.2 and also nearer the end of this thesis in section 10.5 which deals with areas for future research.

Sinclair's work on collocation, in contrast to the textlinguistic dimension investigated by Hoey, is more concerned with relationships between words over shorter distances. Sinclair proposes that any point between words in a text can be considered to be either part of a collocational relationship, or not. Sometimes, at points between words in a text, the choice of the next word is largely open. The lexicon, once it has satisfied local constraints, can provide virtually any word. At other times the language user's choice is restricted to "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair 1991: 110). Text is interpreted according to either of these 'open choice' and 'idiom' principles, but not at the same time. The idiom principle is "as least as important as grammar in the explanation of how meaning arises in text" (ibid.: 112).

One attempt to estimate the relative quantities of these choices found that in one written text, on average, only 45 of every hundred words were single-word choices made, suggesting that "the production of a text involves frequent alternation between prefab and non-prefabricated strings" (Erman & Warren 2000: 51). In addition to writers, speakers have also been found to pause rarely between the component parts of prefabricated strings, suggesting that prefabs "are stored in long-term memory as units" (Erman 2007: 49); this suggests a role for fixed collocational

patterns in cognition. The connection of the idiom principle to Firthian collocation is clear: the meaning of words depends on the other words around them in the text and there is a strong tendency for some combinations of words to co-occur repeatedly.

This closer proximity of a word with its collocate along the syntagmatic plane, Halliday's 'n removes', is referred to as the 'span' by Sinclair (1991: 175), who estimates it as being approximately four words to the left or right of the word. A slightly wider criterion was applied to the COBUILD Collocations dictionary (COBUILD 1995) when the definition of a collocate used was a lexical item "occurring within five words either way of the headword with a greater frequency than the law of averages would lead you to suspect" (Krishnamurthy 1987: 70). Later research by Mason (1997) has questioned whether a span should be symmetrical, i.e. the same number of words to the left or right of the node. It has instead been argued that different nodes have different spans, since "each word form has an individual influence on the variability of its environment" (Mason 1997: 361). This influence is what Mason calls 'lexical gravity', and results in a unique span for any word, based on the restriction the word imposes on the variability of its context.

As work on collocation in the Empirical Firthian tradition has progressed, the single word has become seen as less of a unit of meaning in its own right and more as a constituent of longer meaningful collocations. Hunston and Francis (2000) propose the construct of 'pattern', such as **N n** or **link verb V that** (Hunston 2002: 139). In a more extreme interpretation of Firthian collocation, they argue that "words have no meaning in isolation, or at least are very ambiguous, but have meaning when they occur in a particular phraseology" (Hunston & Francis 2000: 270). Multi-sense words tend to associate each sense most frequently with a different set of patterns, and words with the same pattern tend to share aspects of meaning. Meaning thus belongs "to the whole phrase rather than to individual words in it" (Hunston 2002: 140). A pattern is a compromise between an instance of specific behaviour of a lexical item and a generalisation for which the specific examples are evidence (Hunston & Francis 2000: 81), and "each distinct meaning in a language can be associated with a word pattern that is unique to it" (Sinclair 2004: 281).

This approach to meaning was incorporated into empirically produced reference materials for English language learners. The COBUILD Advanced Learner's Dictionary (Sinclair 2006) has a column containing the pattern of a word used for a particular sense, and the two books on grammar patterns (Hunston & Francis 1996a, 1996b) contain a list of patterns and the most common verbs, nouns or adjectives used in them. Sinclair envisages a time when

> each meaning will be associated with a phrase that reliably creates that meaning and no other; questions like "What preposition goes with that verb?" will disappear because the most appropriate preposition will be part of the model expression.
>
> (Sinclair 2004: 281).

The fixed collocational patterns in the present study can be viewed as the "model expressions" to which Sinclair refers. They are less abstract than grammar patterns, since their forms are fixed and non-generalisable, and their discourse function is more explicit because it is part of their definition.

Work on extended empirical collocations in the Empirical Firthian tradition, such as Stubbs' construct of the *extended lexical unit* (Stubbs 2001: 87), view the phenomenon in terms of the four levels of collocation, namely the lexical, grammatical, the semantic, and the pragmatic. At the lexical level this involves observing the frequent collocation of individual word-forms or lemmas, while at the syntactic level this involves observing the frequent colligation of word classes. The semantic level is shown by the observed semantic preference, such as lexical sets, semantically related word-forms or lemmas, while the pragmatic aspect involves the discourse prosody: the speaker's attitude and the item's discourse function. In Stubbs' exposition of the model he adds an additional criterion of "distribution in text types", or genre, illustrating this with the behaviour of the lemma UNDERGO in "general English" (in the form of a normative sample of the British National Corpus (see section 4.3.1 below)). In its left-hand environment, UNDERGO has an "involuntary" discourse prosody and in its right-hand environment a discourse prosody of "unpleasantness": people are forced to undergo unpleasant experiences. In technical English data, by contrast, Stubbs finds that it has "no necessarily unpleasant connotations" (ibid.: 91).

The two different perspectives on collocation, the Eastern European and the Empirical Firthian traditions, can be applied to different collocational units using similar criteria, as shown in Table 2.2 below:

Criterion/Level	Eastern European Perspective	Empirical Firthian Perspective
Lexical	Limited commutability	Frequent collocation of individual word-forms or lemmas
Syntactic	Grammatically well-formed (an adverb modifying an adjective);	Observed frequent colligation of word classes
Semantic	Partially semantically transparent, in which one element of the combination has a literal, unidiomatic meaning	Observed semantic preference, such as lexical sets, semantically related word-forms or lemmas
Pragmatic	Institutionalisation; distinctive and memorised	Discourse prosody; descriptor of speaker attitude and discourse function

Table 2.2: Comparison of Eastern European and Empirical Firthian approaches to collocation

Little investigation has yet been conducted into the collocational environments of fixed collocational patterns along these lines. These four levels, together with the wider level of genre, constitute a heuristic framework with which to investigate fixed collocational patterns. At the lexical level their form is fixed, and so they are not commutable according to the Eastern European perspective. A fixed collocation with another word substituted would become another fixed collocational pattern, provided it was seen to occur frequently enough. From the Empirical Firthian perspective they are frequent collocations of individual word forms.

The three other criteria, together with Stubbs' wider criterion of text type (or genre), can be used to investigate the use of fixed collocational patterns in research articles:

- At the syntactic level the study will investigate whether there is any relation between the position of the combination in the sentence and its adjacent lexis, its location in the text, and its role in the discourse.
- At the semantic level, is there any relation between the lexis adjacent to the pattern and its role in the discourse?

- In terms of function, is there any relation between the pattern's location in the text and its role in the discourse, e.g. does a combination occurring in a named section, such as the results/discussion section, perform a function associated with conclusions?

Textual positioning can be determined by use of a discourse map (Biber et al. 1998: 122), on which the location of occurrences of a linguistic feature is plotted through the text. In this case the location of each pattern in the text in which it occurs can be determined and compared to its observed function.

We have seen that work on collocation began with studying the influence on the semantics of a single word by its lexical relations with other words, and subsequently took into account the influence of other factors such as pragmatics. The next section focuses on collocations which consist of longer strings of words.

2.3.2 Extended empirical collocations

The fixed collocational patterns investigated in the present study are extended collocations, recurring fixed strings of three or more words. In computational linguistics these are referred to as "n-grams," strings of *n* words where *n* is greater than one. Table 2.3 below shows some examples of 4-grams identified in the *Google Public Web Training Corpus* of 1.025 trillion tokens (Franz & Brants 2006). This is a vast amount of data compared to the data sets on which previous empirical investigations of collocations were conducted.

4-gram	Frequency
serve as the incoming	92
serve as the incubator	99
serve as the independent	794
serve as the index	223
serve as the indication	72
serve as the indicator	120
serve as the indicators	45
serve as the indispensable	111
serve as the indispensible	40
serve as the individual	234

Table 2.3: Examples of 4-grams from the Google Public Web Training Corpus

N-grams reflect the computer-mediated view of language as a sequence of characters and spaces, and so such strings do not follow traditional grammatical boundaries. They are identified regardless of their meaning, use, or the word classes their components might belong to. It can be seen that the above 4-grams only partially correlate with traditionally recognised linguistic structures. They contain an intransitive verb and a fragment of a prepositional phrase: *serve as the* ____. The prepositional phrase in turn contains a noun group such as *the indication*, *the index*, or a fragment of a longer noun group such as *the incoming*, or *the indispensable*. These n-grams are identified without regard to any mis-spelling: *serve as the indispensable* and *serve as the indispensible* are separate n-grams. The methodology for obtaining n-grams has been used in applications of Corpus Linguistics to the study of different spoken and English registers, particularly in the case of the *lexical bundles* described by Biber et al. (1999). These are the principal items with which the results from the present study will be compared.

The theoretical and methodological aspects of lexical bundles overlap with Phraseology, Academic Discourse, and Corpus Linguistics, and so they will be discussed in relation to each field in different parts of this thesis: here in this section, in section 3.3, and in section 4.4.3. Lexical bundles are frequently recurring extended collocations of three or more words which show "a statistical tendency to co-occur" (Biber et al. 1999: 989) and therefore "commonly go together in natural discourse" (ibid.: 990). Lexical bundles are corpus-driven phenomena and are used by Biber as an indicator of register. They can be spoken, as in *going to be a* and *I said to him*, or written, as in *in contrast to the*, *one of the*, *the use of*, and *on the other hand.*

The statistical measure used by Biber to show significance is, as in the present study, absolute frequency. The frequency criterion specified, by which an extended collocation can be deemed a lexical bundle, is more than ten times per million words for 4-word bundles, whereas for 5- and 6-word bundles it is at least five times per million words (ibid.: 993). In addition, lexical bundles must occur in at least five texts in the register so that idiosyncratic uses by individual authors do not bias the study. One difference between n-grams and lexical bundles is that lexical bundles cannot be interrupted by punctuation, or sentence or turn boundaries. By definition, shorter lexical bundles are more frequent than longer bundles, since the former occur both in

their own right and within the longer bundles. While, for example, the 4-word lexical bundle *the use of the* occurs 2,491 times in the Bank of English, the 3- word bundle *the use of* occurs 21,094 times, which includes its 2,491 occurrences within *the use of the*.

The lexical bundle appears to be still in development as a construct. The *lexical* part of the term is slightly problematic because the empirical nature of the identification criteria means that there is no guarantee that such frequently occurring fixed strings will always contain a "lexical" or "open class" word. Many extended collocations such as *and in off the* or *that and the* consist of entirely grammatical or closed class words. While these extended collocations are fairly infrequent on the Bank of English and the British National Corpus, they could conceivably occur more than 10 times per million words in larger corpora such as the Google Public Web Training Corpus. The frequency criterion is also arbitrary: in later work on lexical bundles (Cortes 2002; Biber, Conrad, & Cortes 2004; Cortes 2004; Biber & Barbieri 2007) this is increased to 40 occurrences per million.

The patterns to be investigated in this study have the same form as lexical bundles but are not identified using the same criteria, hence the term *fixed collocation pattern* is used as an umbrella term to include extended empirical collocations such as lexical bundles and other similar forms. The functional framework for lexical bundles developed by Biber and Hyland will be discussed in section 3.3 below.

The other fixed collocation pattern that will be compared with the results of the present study is the "written academic formula" found by Ellis and Simpson (2005). Their study aimed at finding psychologically salient fixed collocational patterns for learner writers of academic English. They used a very different methodology from that of the present study: a corpus of 9 million tokens was collected comprising various publicly available written and spoken academic and non-academic English corpora, and their formula candidates were retrieved from the data via *Collocate* software (Barlow 2004a). From these candidates they identified written academic formulas using a triangular identification procedure, matching corpus validity, educational validity, and psychological validity. To establish corpus validity, log likelihood ratios were used to identify combinations which were significantly more

common in academic writing than in other registers of English. To establish educational validity, teachers of academic English were asked to rate each formula according to whether they thought it was a fixed phrase, had a cohesive meaning or function, and was seen as worth teaching. To establish psychological validity, phrases and non-phrases were presented as visual stimuli to native speakers who rated them in terms of their acceptability. Table 2.4 below lists the top 20 written academic formulas in their study:

Written Academic Formula	**Weighted z score**
as well as	3.27
in order to	2.96
the use of	2.81
the number of	2.55
there is a	2.43
there is no	2.32
a number of	2.32
on the other hand	2.26
it is not	1.99
that it is	1.59
in the case of	1.59
the other hand	1.49
the case of	1.47
as a result	1.41
based on the	1.38
on the other	1.36
in the case	1.34
in which the	1.29
likely to be	1.24
on the basis of	1.23

Table 2.4: Top 20 written academic formulas (Ellis & Simpson 2005)

These written academic formulas are fixed collocational patterns identified from different corpus data using statistical measures not used in the present study, and so will be a useful alternative benchmark for comparison with the fixed collocational patterns to be identified in the present study. Having examined the two main empirical approaches to collocation, we now turn to the third aspect of word combinations, that of the context of situation in which they are used.

2.4 The situational contexts of word combinations in use

The third area of word combinations of relevance to this study, and which I have already touched on when referring to register and genre, involves the situation in which word combinations are used. Much work has been done on word combinations, spoken or written, which perform some function in a social interaction. Many social interactions, such as when someone greets a friend, or a lecturer gives a presentation in an airless lecture theatre, or a diplomat concludes a polite exchange with a host government, take place in physical and social settings which are familiar to the participants. Participants in a social interaction each have been in similar situations many times before, and will have a good idea of what the other participant expects to happen during the exchange, and therefore will know the linguistic choices required for the completion of the interaction in an appropriate way. The language used in common social interactions consequently tends to consist of a high proportion of well-worn, rather than newly-minted and original, extended collocations, e.g. *How are you?; Could somebody open a window please?; I have the honour to convey to Your Excellency the assurance of my highest consideration.*

To many of the researchers in this area, when users use the same word combination over and over again its character, if not its form, becomes different from language which has been generated on a one-off basis for a unique situation with which the participants are not familiar. Word combinations which are regularly used in specific communicative situations are often seen as "formulaic" utterances made for social reasons rather than to achieve any specific communicative goal. Wray (2002) focuses on this aspect of collocation in her construct of the *formulaic sequence*, a very flexibly-defined continuous or discontinuous collocation which is highly context-dependent and pragmatically very efficient.

2.4.1 Word combinations in social situations: creativity and cliché

A particular extended collocation or formulaic sequence is "pragmatically efficient" if it is what is understood to be what is normally said in the circumstances: in many social situations there are phrases which are seen as being particularly appropriate

to the occasion. Knowledge of the collocations which make up these phrases therefore comprises part of a language user's communicative competence (Hymes 1971). Through repeated exposure, use, and feedback, the collocations used in successful communication in these situations are internalised by language users, so that in effect these collocations constitute a single choice, a formula which the user knows will contribute to the success of a social interaction.

One example of a highly specialised social interaction involving a relatively fixed extended collocation is in a diplomatic exchange between two countries. At the end of a diplomatic Note, the formula shown in Example 2.4 is always used:

> Example 2.4: The Embassy of [country A] avails itself of this opportunity to renew to the Ministry of Foreign Affairs of [country B] the assurances of its highest consideration.

Apart from the variation permitted with the insertion of the names of the countries involved in the exchange, this phrase is completely fixed: these words are always used in exactly this order and "lend a degree of gravity to the content that would perhaps be missing if the same message was passed in a letter" (Protocol Division of the Foreign and Commonwealth Office, Personal Communication). At this point in the communication there is little room for any linguistic creativity on the part of the diplomat writing the Note. Regardless of its actual contents, the omission of this formula by the writer, or the slightest variation – intentional or not – from the collocations of which it is composed, could have serious diplomatic consequences. The use of extended collocations in certain communicative situations is mandated by the weight of convention and precedent established by previous users of the language, and by the expectations of current members of the discourse community.

However, while empirical linguists observe and describe the role of collocations in particular communicative situations, language users themselves often make value judgements about extended collocations which appear to be neither creative nor original. Some extended collocations are seen as *clichés,* which are felt to have lost their force through over-use (Howarth 1996: 13). Where the empirical linguist observes "a large number of semi-preconstructed phrases that constitute single choices" (Sinclair 1991: 110), the language user sees "phrases tacked together like the sections of a prefabricated hen-house" (Orwell 1946: 3; Parris 2007). The Irish

writer Flann O'Brien (writing as 'Myles na Gopaleen') pokes fun at this kind of over-use in his *catechism of cliché*, a "unique compendium which is all that is nauseating in contemporary writing" (O'Brien 1968: 202):

> If a thing is fraught, with what is it fraught?
> The gravest consequences.
> What does one sometimes have it on?
> The most unimpeachable authority.
> (ibid.: 209-210)

So, while using fixed collocational patterns such as *fraught with the gravest consequences* and *on the most unimpeachable authority* may not always be the height of creativity, and thus fail to count as good language style, in certain situations their employment is mandated by the weight of convention and precedent established by previous users of the language, and also by the expectations of current members of the discourse community.

There is consequently an apparent tension between creativity and collocation in certain communicative situations:

> what we value most in language - creativity, expressiveness [...] - allows us to succeed less well in having others understand us than the largely prefabricated phrases we use to say almost the same thing over and over again. Paradoxically, language is at its best when it matters least; at its worst when it matters most.
> (Moore & Carling 1988: 72)

There is a difference between *perceived* and *observed* frequent use, however. A collocation which appears to be used *repetitively* may not in fact be used *repeatedly*, and *vice versa*. As mentioned earlier in this chapter, empirical work on collocation has revealed that clichés and idioms are rare in the use of English as a whole (Moon 1998), and that they only *appear* to be over-used by speakers in particular domains or genres. The fixed collocational patterns in this study are certainly used repeatedly but do not immediately appear clichéd.

The tension between the creativity of the language user and the urge to conform to expectations by using a restricted number of collocations is very apparent in the context of English literature, when writers deliberately try to break collocational relations between words in order to say something new:

> the prose writer or the poet strives, over a larger or smaller stretch of text, to convey something which he cannot achieve by normal means, and he thereby sets us a problem in which we can lean on no experience of directly relevant instances.
> (McIntosh 1961: 336)

The communicative purpose of the writer of a novel is situated in the literary genre in which they are writing, and literary authors can go to great lengths to avoid using language which appears to them to be unoriginal and lacking creativity. Smith (2007), for example, is highly self-critical when she describes occasions when she has been unable to break collocational relations successfully:

> ... in each of my novels somebody "rummages in their purse" for something because I was too lazy and thoughtless and unawake to separate "purse" from its old, persistent friend "rummage". To rummage through a purse is to sleepwalk through a sentence - a small enough betrayal of self, but a betrayal all the same.
> (Smith 2007: 3)

Figure 2.1 below shows evidence from the Bank of English of the collocates of the noun *purse*. Without distinguishing between the British sense of *purse* (in American English a *change purse*) or the American sense (in British English a *handbag*), it can be seen that *rummage* is the second most frequent collocate after *fumble*. To find a more original collocation in an attempt to convey a more profound literary meaning, the writer would need to choose a verb with a conceivably similar meaning to *rummage*, *fumble* and so on, but which the reader is unlikely to have encountered before in connection with *purse*.

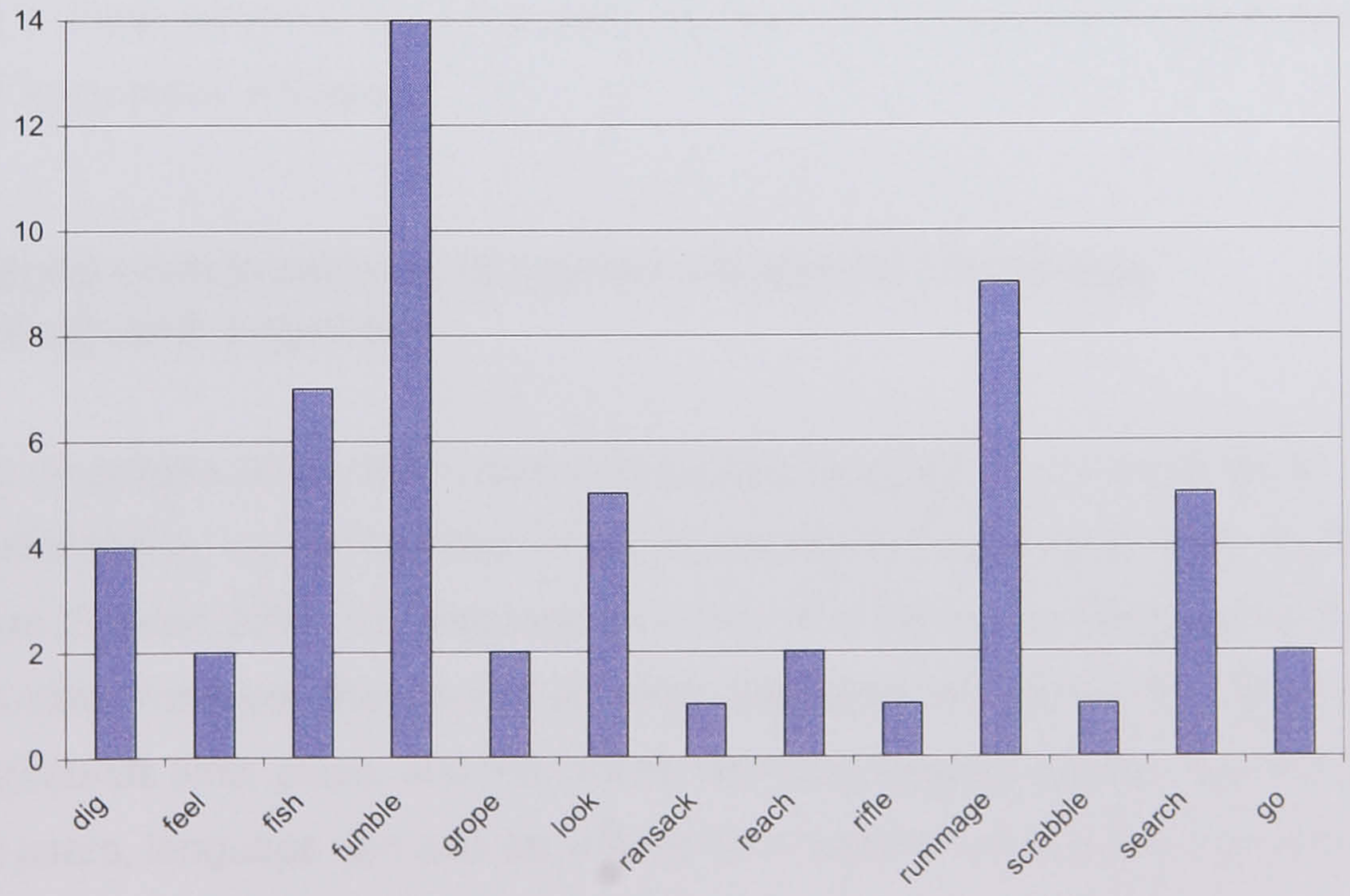


Figure 2.1: Frequent left-hand verb collocates of the noun *purse* in the Bank of English

Extended collocations of various kinds have been seen as the means by which users conform to the communicative norms of a particular discourse community in order to bid for inclusion in that community. Carter (2004) has pointed out the interplay between the individual language user's conventional and appropriate use of these forms and their creativity in communicating with others in the context of a particular social situation. Creativity in this sense, however, is not merely the "breaking or bending of rules and norms of language, including a deliberate play with its forms and its potential for meaning," (ibid.: 9) like the innovative uses of words aimed at by writers in literary genres such as Smith. It also has an important interpersonal function in spoken communicative situations.

The communicative situation in which the fixed collocational patterns in this study will be investigated is the genre of the research article in different disciplines. It will be seen below that research article lengths in different disciplines vary so widely that it is difficult to collect a comparable amount of data in each discipline. Consequently, one of the main questions of this study will be to find out whether texts are a better measure of comparison than tokens. Are similar fixed collocational patterns found in a comparison between equal amounts of language, i.e. an isolexical comparison, and in a comparison between an equal number of communicative situations, i.e. an isotextual comparison? If the same patterns are frequent in both comparisons, then the findings from the isotextual comparison should reveal more about the pragmatic functions of these patterns, since the same number of communicative acts in each discipline have been compared.

2.4.2 Word combinations in social situations: language processing and cognition

The focus on written data in this thesis investigates language as an observable social phenomenon, one instantiated in social practices in the form of texts, or, to paraphrase (Sinclair 2001: ix), language on paper or in the air, not language in the mind. It is clear, however, that as well as a set of systems of symbols and meanings which transcends time, place, and the mental and physiological capabilities of its individual users, language can also be viewed from another perspective as a mental and psychological phenomenon used by individuals to communicate in social

situations. The role of word combinations in cognition is outside the main scope of this study, but some attempts have been made to look at the psycholinguistic validity of fixed collocational patterns from corpus data, e.g. Schmitt et al. (2004) which will be returned to in section 10.5, and so there is a brief discussion here of the main areas in relation to artificial intelligence and cognitive psychology.

In the 1970s Bolinger focused on word combinations as a way of providing an alternative to the dominant Chomskian perspective on language of generally applicable rules, a lexicon of minimal units, and a set of semantic principles of orientation (Cowie 1998a: 2). The claim that fixed forms are commonly observed in language use because users have learned, and in someway stored mentally, these forms was initially contentious, as it suggests that individuals gain part of their mental language knowledge from interaction with the semiotic system of the speech community, rather than wholly relying on a biological language faculty. Language "learning goes on constantly ... in segments of collocation size as much as it does in segments of word size, and ... much if not most of our later manipulative grasp of words is by way of analysis of collocations" (Bolinger 1976: 8). The alternative to the prevailing Chomskian view was that speakers possess a non-homogeneous store of language knowledge, consisting of the set of generative rules and also a store of pre-assembled patterns. But a speaker may at times 'bypass' the system of grammatical rules and retrieve a word combination instead:

> a good stock of the wordings we use are stored a higher rank, ranging from formulaic expressions like *the manager will see you in a minute* through *it needs to be put on a sound commercial footing* to the small change of family life like *have you remembered to take your Vitamin C?* and *where's that cat?.* Proverbial sayings provide an extreme case of learned syntagms stored at higher rank, but they are by no means unique."
>
> (Halliday 1994: xxi)

Similar conclusions were drawn in the field of artificial intelligence, where attempts to model language for comprehension and production by computer algorithms were impeded by the lack of Chomskian interest in language use:

> the "modern" linguist spends his or her time starring or unstarring terse unlikely sentences like "John, Bill, and Tom killed each other" (to pick one at random from a recent journal), which seethe with repressed frustration and are difficult to work into a conversation.
>
> (Becker 1975: 1)

As a computational linguist whose goal was to replicate human reception, processing, understanding, and production of language, Becker felt that any theory of language which favoured memory storage over processing power became easier to model with the available computer technology at the time. Becker proposed a view of language whereby "utterances are composed by the recitation, modification, concatenation, and interdigitation of previously-known phrases consisting of more than one word" (ibid.). In other words, language users mostly put together pieces of text that they have heard before, and productive processes have the secondary role of adapting the old phrases to the new situation, i.e. the concatenation, modification and so on referred to.

Parallels can be drawn between the work of Becker and research in cognitive psychology, which, by analysing spoken and to a lesser extent written language in use, attempts to model the nature of language knowledge in the mind: how it is stored, processed, retrieved for production, and recognised, decoded and understood. One of the key papers in this area, and which can be found in the literature reviews of analyses of word combinations from virtually any theoretical perspective, is that of Pawley and Syder (1983). They made a similar point to that of Becker, arguing that language users do not fully exploit the creative potential of syntactic rules, and "only a small proportion of the total set of grammatical sentences are nativelike in form" (Pawley & Syder 1983: 171). They defined 'nativelike in form' sentences as those which are "ordinary and natural, rather than 'unidiomatic', 'odd' or 'foreignisms'" (ibid.).

In an attempt to account for this nativelike selection and nativelike fluency, Pawley and Syder hypothesised that a language user has a store of memorised sentences and 'lexicalised sentence stems' which "the speaker is capable of consciously assembling or analysing, but which on most occasions are recalled as wholes or as automatically chained strings" (ibid.: 205). Only "a minority of spoken clauses are entirely novel creations in the sense that the combination of lexical items is new to the speaker" (ibid.). Fluency is used more in familiar situations, while creative, less fluent, language is used in novel or unexpected situations. These findings are still widely quoted in many subsequent papers on the phenomenon, although they were based on a relatively small amount of conversation data. The situated nature of such

aids to fluency has obvious parallels to work on word combinations in different social settings discussed in section 2.4.1. The travel writer Stuart Maconie amusingly describes the effect of a sudden unexpected change in the situation on a language user's performance, when familiar word combinations can no longer be used and the user's fluent production becomes dysfluent. He observes a senior conductor making announcements on a train who "declaim[s] with the verbiage and prolixity of Laurence Olivier doing Shakespeare, until something goes wrong, when he suddenly turns into Buster Keaton doing Harold Pinter, all mysterious pauses and menacing silences" (2007: 35).

Wray and Perkins seek to highlight the role of formulaicity in language processing and production, with creativity complementing the process, again in a similar way to that suggested by Becker:

> the best deal in communicative language processing is achieved by a suitable balance between creative and holistic processes. The advantage of the creative system is the freedom to produce or decode the unexpected. The advantage of the holistic system is economy of effort when dealing with the expected."
>
> (Wray & Perkins 2000: 11).

Utterances can be novel in that they are "tailor made to a particular situation, but at the same time can conform to a paradigm of sequences" (ibid.: 12). Much of the work cited in this section does not rely on corpus evidence, and relatively few studies have attempted so far to evaluate these claims, with the exception of those by Ellis and Simpson (2005) and Erman (2007). Another study by Schmitt et al. (2004) was less successful in establishing a link between fixed collocational patterns and cognition. The study by Schmitt et al. will be discussed near the end of this thesis in section 10.5 as a basis for further research.

2.4.3 Word combinations and English as a lingua franca

A final consideration relating to the fixed collocational patterns in this study is that of the role of the native speaker. The approaches to collocation outlined above rely on the English native speaker as a data source: intuitive descriptions of collocational relations rely on a native speaker's judgement, while corpora such as the Brown or

the British National Corpus (see section 4.3.1 below) have been designed to represent the language used by American or British native speakers, and so the collocational tendencies identified normally reflect native speaker usage. These approaches to collocation are being re-assessed in the light of the unique status of English as both a language with different standard versions around the world (Kachru 1992) and as a *lingua franca* in which the majority of communication worldwide now involves one or more non-native speakers (Crystal 2003).

The main area of uncertainty is the place of native speaker varieties of English in English language teaching and learning. If a minority of the users of a language speak it as a mother tongue, then it is not clear which variety of the language should continue to be taught to its learners. The variety of English chosen as the model for learning will ultimately depend on the communicative situations in which learners intend, or will be obliged, to use it. Some learners will prefer to learn English the way native speakers use it (Timmis 2002) since "being perceived as a member of a certain linguistic group that speaks the L2 natively ... is also important to certain learners of a language" (Nesselhauf 2005: 2). On the other hand, some learners might be suspicious of the ideological assumptions behind a native speaker model. Over the past two decades, several writers (Phillipson 1988; Pennycook 1994; Canagarajah 1999) have warned that politically and economically powerful native English-speaking countries seek to perpetuate and extend their dominance by disseminating a native English-speaking view of the world through the "imperial troopers" (Edge 2003: 10) of the global English language teaching industry.

In the field of Academic Discourse, it may be difficult to predict the possible effects of releasing English, and thereby its collocations, from the obligation to be acceptable to native speakers. Johns (1996: 1) identifies a "slight oddity" arising from a learner's collocation of *present* and *insight* in Example 2.5:

> Example 2.5: The variation of these ratios presents some insight about the financial intermediary role of banks.

After referring to a corpus of technical English, Johns suggests that the sentence could be improved using *present* with its more frequent collocate *findings*, or by using *offer* with its frequent collocate *insights*. Users of English as a lingua franca may come to have higher tolerances of formerly odd-sounding collocations, since by

definition these will no longer sound "odd". Alternatively, lingua franca users might insist upon continued restrictions on collocability. Defining a suitable lingua franca speaker is not straightforward, however (Prodromou 2006). Initial moves towards identifying novel and continuing tendencies of words to co-occur have already been made with the VOICE corpus (Seidlhofer 2001). Descriptions of lingua franca collocational patterning have also been reported (Mauranen 2004) which suggest that non-native speaker use continues at times to approximate to native speaker use.

The corpus I have collected for the study of fixed collocational patterns does not aim to represent native, non-native, or lingua franca speaker use. The status of the writer of the text is not relevant to the decision whether or not to include a text in the corpus: a text only needs to have satisfied the expectations of the discourse community in each discipline by being published in a journal considered to be prestigious by other members of the discourse community (see Appendix 1). It can then be considered to belong to the genre of the research article (see section 3.2.3 below).

It is highly likely that a large number of texts in the corpus collected for this study will have been written by lingua franca speakers, given that academic discourse communities are global in scope. Some "oddities" will as a result be observed in the collocational behaviour of particular items in the corpus. For example, the usual **V n** pattern of the reporting verb *obtain* is sometimes extended (Hunston 2007 using Oakey's data) by lingua franca speakers to fit the **V that** pattern of other reporting verbs such as *find*, *claim* or *argue*, as shown in the three examples below:

> Example 2.6: In the case of the QD dimension with the lateral size of 10 nm and the height of 3 nm, which is the case of our measured QD, we have obtained that the biexciton binding energy is negative and the magnitude is 27.1 meV.

> Example 2.7: We concluded that the collisionless dissipation occurs and we obtained that the damping time is about 40 fm/c for the energy density (160 MeV)4.

> Example 2.8: In what respects to citizen participation, it was obtained that each and every dimension of market orientation has a significant relationship with it, confirming H7a (partially), HA, and H7c.

These examples are all used at a particular stage of a successful communicative act, and so their correctness or incorrectness is a moot point.

2.5 Summary

This chapter has attempted to establish the fixed collocational pattern in relation to studies of textual meaning, empirical studies of collocation, and its use in particular communicative situations. A framework of four criteria, derived from the Eastern European and Empirical Firthian approaches, has been proposed as a possible heuristic for further investigating the role of fixed collocational patterns in the genre of the research article, namely by examining their lexical, syntactic, semantic, and pragmatic levels. The thesis now moves on to examine the specific context of situation in which the fixed collocational patterns in this study will be investigated: the research article.

Chapter 3 Academic English discourse

3.1 Introduction

So far in this thesis I have argued that fixed collocational patterns are extended empirical collocations grounded in the Empirical Firthian approach to meaning as context of situation. We now move on to the situation in which the fixed collocational patterns in this study are described, namely the research article. This chapter surveys the different approaches to the description of the linguistic features that differentiate research articles from other written texts. It will discuss how comparisons between varieties can be made at different levels, such as *text*, *genre*, or *register*, and briefly surveys the different theoretical perspectives in which these terms are used.

3.1.1 Discourse analysis

Descriptions of written academic English belong to a long tradition of discourse analysis (Sinclair & Coulthard 1975; Stubbs 1983; Schiffrin 1993; J. Flowerdew 2002). Much of this work has been concerned not only with analysis of linguistic forms, but also with the wider social practices and contexts in which the discourse is produced (Myers 1990; Hyland 2000). Through looking at records of instances of spoken or written human communication, more can be learned about the social semiotic systems in and through which the communication takes place, and so more can be learned about the relationship between culture, society, meaning and language. This focus on language and social interaction makes discourse analysis an area of Sociolinguistics, although as a methodology it is also used in Psychology, Anthropology, Sociology, Philosophy, and Artificial Intelligence (Stubbs 1983: 12).

The primary unit of discourse analysis is the communicative act: discourse analysts work with spoken or written texts, and stretches of language "above the sentence or above the clause" (ibid.). One aim of discourse analysis is typological: describing, categorising, and accounting for the language used in social interactions to reveal the relationships between the participants and into which they can be classed as being distinct from other instances of communication (McCarthy 2001). Discourse

analyses have been performed in a wide range of social interactions, such as service encounters (Koester 2004: 88), radio phone-ins (McCarthy & O'Keeffe 2003), business meetings (Handford & McCarthy 2004), as well as the academic journal articles to be used in this study.

In studies of Academic Discourse, the primary typological goals have been to classify texts and other instances of communication in different disciplines (J. Flowerdew 2002). These typologies have had different goals: in Anthropology, the aim of this grouping is to make clearer the epistemological differences between disciplines, namely the way each discipline makes knowledge. In Education and the pedagogical field of English for Specific Purposes (ESP), the goal is to match linguistic descriptions to pedagogical needs of learners of academic English who will be expected to use it in similar situations. In Corpus Linguistics the goal is to allow contrastive corpora to be collected so that comparative studies of linguistic features can be conducted.

3.2 Descriptions of written academic discourse

The study of academic discourse for pedagogic purposes intensified during the second half of the 20th century, when English became widespread as the global language of communication in the fields of Science, Technology, Business and Academia (Phillipson 1988; Pennycook 1994; Dudley-Evans & St John 1998: 19; Crystal 2003). There was a corresponding increase in the number of learners of English who needed to use it to communicate in these settings, and a lack of suitable materials from which they could learn. This need for teaching materials contributed to the beginnings of the field of applied linguistics when, in *The Linguistic Sciences and Language Teaching*, Halliday et al. (1964) argued that materials for these new learners should be based on sound linguistic descriptions of the language used in these situations: "detailed studies of restricted language and special registers [should be] carried out on the basis of large samples of the language used by the particular persons concerned" (1964: 190). Before linguistic analyses could be performed, however, there was a need to develop criteria for establishing texts as belonging to a particular type, or belonging to a particular register, or being instances of a particular genre, which could then be grouped together and studied.

Register and *genre* are somewhat contested terms: they have specific roles in Systemic Linguistics, while in Corpus and Applied Linguistics they are more related to discourse typology. The next sections will explore their possible meanings further with respect to theoretical models and explain how they will be used in this study.

3.2.1 Hallidayan approaches to Language, Register, and Genre in Systemic Functional Linguistics

The concepts of language, register, and genre in Halliday's systemic-functional model of language (Halliday 1973, 1978, 1994) were developed from Firth's polysystemic theory of meaning, and so, like Sinclair's work on collocation described in Chapter 2, derive from the anthropological work of Malinowski in the 1920s. Halliday, and later Martin, formulated the systemic relationship between language and Malinowski's notions of meaning as context of situation and context of culture, eventually developing the model now used in the field of Systemic Functional Linguistics. In this model, language is not only a semiotic system itself, but is also the means by which meaning is expressed through two other semiotic systems, namely register and genre. The term *register* in this model refers to the systematic relationship between language and Malinowski's context of situation, while *genre* corresponds to Malinowski's broader notion of context of culture (Martin 2001: 152). Genre acts on register by constraining the possible combinations of different categories of register, and register in turn constrains language by limiting the linguistic choices available to the participants.

Register in this model consists of three categories: the *field* of discourse, i.e. the area of operation of language activity; the *mode* of discourse, primarily spoken or written, and the *tenor* of discourse, the relationship between discourse participants. The relationship between language and meaning as context of situation is the correspondence of these three categories of register to three aspects of meaning, termed metafunctions: *field* corresponds to ideational meaning; *mode* relates to textual meaning; *tenor* to interpersonal meaning (Halliday 1994: xiii; Martin 2001: 154). The most abstract semiotic system, genre, is a "staged, goal-oriented purposeful activity in which speakers engage as members of our culture" (Martin 2001: 155). This framework is illustrated in tabular form in Figure 3.1 below:

Context of culture	Genre		
Context of situation	Register		
	Field	Tenor	Mode
Functional component	Ideational meaning	Interpersonal meaning	Textual meaning

Figure 3.1: Martin's framework of genre and register related to the contexts of culture and situation.

The systemic functional framework does not only seek to describe how language performs these functions with the other semiotic systems, but also attempts to account for how "the form of language has evolved to serve these functions and in the process been shaped by the need to perform these functions" (Halliday 1973: 7). For example, once a language develops a writing system, the structure of that language is influenced by its use in writing (Martin 2001: 153). While studies of academic discourse do not usually use the whole of Halliday and Martin's systemic functional framework, their three functional components have been employed as the basis for analyses of the roles of extended collocations in academic discourse. These will be further discussed in section 3.3 below.

3.2.2 Register Analysis

A different framework for describing academic discourse, which somewhat confusingly uses the terms *register* and *genre* in another way, is employed in the field of ESP. While genre encompasses register in the Hallidayan approach outlined above, register in the ESP perspective is a broader category than genre, and refers to language used in a particular situation. The term is "used in a less theoretically precise way to mean simply discourse occurring in a particular context" (Hunston 2002: 160). Early in the history of ESP, texts belonging to the academic register were seen as easier to analyse scientifically than the literary texts which had previously been the object of linguistic analysis (Swales 2001). Academic texts could be described rationally and objectively because, instead of being polluted by literary

or stylistic pretensions, they were "rhetorically simple and transparent linguistic mechanisms for the display and transmission of knowledge, hypotheses, methods and experimental results" (ibid.: 43).

These early analyses of academic register aimed to describe how the linguistic forms in texts in one register differed from those in other registers. Barber (1962), for example, focused on quantitative studies of different linguistic features, such as the ratio between the number of types (individual word forms) to the number of tokens (total word forms) in academic texts. Texts from the Life Sciences were found to have higher type/token ratios than other texts, which meant that they had fewer repeated words, and thus were more lexically "dense". Other features examined were the number of simple sentences, the number of passives, and so on. This type of approach was useful for identifying academic texts but had little to say about why such features were common, i.e. why academic writers viewed the use of large numbers of passives as appropriate in that context of situation. The main focus in the 1970s moved to relating language form to language use within texts, and resulted in the analysis of sections of texts in academic register in terms of their rhetorical functions, such as definition, classification, description and hypothesising (Dudley-Evans & St John 1998: 23).

Later register analysis has attempted to address these issues through the use of corpus linguistic techniques. Biber et al. (1998) compared the distribution of selected grammatical features across a corpus of different registers and then applied these "to more comprehensive register characterisations" (ibid.: 138). They also undertook a genre analysis (see section 3.2.3 below) by looking at linguistic variation across texts from Biology and History and "how these linguistic patterns are related to the primary purposes and methods of each discipline" (ibid.: 139). Biber's research focus then moved from the investigation of grammatical features across different registers to the investigation of extended empirical collocations, i.e. lexical bundles, across registers. This work will be further discussed in section 3.3 below.

3.2.3 Genre Analysis

In ESP, the field of Genre Analysis (Swales 1990; Bhatia 1993) or Academic Discourse Analysis (J. Flowerdew 2002) treats genre as a class of communicative acts. Writers produce texts for particular purposes, and if the writer of a text belongs to a discourse community whose members also produce texts with similar communicative goals, then these texts constitute a genre. According to Swales (1990: 52), the rationale behind a particular genre is determined by the expert members of the parent discourse community, who set the "constraints on allowable contributions in terms of their content, positioning and form" (ibid.: 52). A writer wishing to produce a text acceptable to the experts in the parent discourse community must therefore pay attention to content, organisation and form in order to ensure his or her text is an allowable contribution. Writers who aim to join this discourse community and need to be able to produce these genres need to learn how to produce these genres.

As a result, much genre analysis has been conducted in order to identify and describe the lexico-grammatical patterning of texts belonging to different genres so that they can be taught to writers wishing to become members of the relevant discourse community (J. Flowerdew 2002). The field of genre analysis accordingly has a narrower focus than register analysis. It looks at specific genres rather than the register of "academic writing" in general. Its focus is also deeper, looking not just at formal features but also at communicative purposes (Swales 1990: 3).

One of the most studied genres of academic writing is the research article, typically those published in peer-reviewed journals. The research article is the primary means by which knowledge is peer-reviewed, assessed and disseminated, although it has been pointed out that another genre, that of the patent, is said to hold more than 80% of mankind's technical knowledge, much of which is not available to researchers in any other form (Groom 2002). Early work on research articles was formal, such as that of West (1980), who examined the occurrence of *that*-nominals. In terms of the communicative purpose of the genre, Hunston (1993) looked at evaluation in scientific research articles, while Myers (1990) compared research articles with popular scientific articles.

Other studies of research articles have also looked at stages of the genre, e.g. article introductions (Swales 1981, 1990) or discussion sections (Dudley-Evans 1986; Hopkins & Dudley-Evans 1988), conclusions (Hewings 1993) or acknowledgements sections (Bloor & Bloor 1989; Hyland 2003, 2004). These studies have revealed the rhetorical moves associated with particular stages of the discourse, and the typical linguistic realisations of these moves. Swales (1990), for example, developed the influential CARS ("create a research space") model of research article introductions (ibid.: 141). He identifies the communicative purposes behind three characteristic 'moves': establishing a territory, establishing a niche and then occupying the niche.

Research on extended collocations in the research article genre has been conducted by Hyland (1996b; 1996a; 1998b; 1998a; 2000) who, in a series of related articles, writes on the role of metadiscourse in structuring texts in different disciplines. Gledhill (1999; Gledhill 2000b, 2000a) has also examined the discourse function of collocations of function words in cancer research articles. As these investigations involve the study of a corpus, they will be discussed in section 4.4 below. Work on the different stages of the genre of the research article has revealed the different discourse functions employed. The next section explores in more detail work on fixed collocational patterns which are seen to perform discourse functions in this genre.

3.3 Functions of fixed collocational patterns in academic English discourse

This section attempts to relate the frequently occurring fixed collocational patterns investigated in this study to the two most recent studies of lexical bundles in academic discourse available, those of Biber (2006) and Hyland (2008). Both these studies provide functional frameworks for lexical bundles observed in register corpora. Both compare lexical bundles in different disciplines: Biber (2006: 166-168) presents a genre-based study of the functions of lexical bundles found in the textbook section of the T2K-SWAL corpus, comprising the disciplines of Business, Engineering, Natural Science, Social Science, and Humanities disciplines (this corpus is discussed in more detail in section 4.4.3 below). Hyland (2008: 13-14), on the other hand, presents a register-based study of the functions of lexical bundles in a corpus of academic prose, containing research articles, dissertations and MA

theses in the disciplines of Biology, Electrical Engineering, Applied Linguistics, and Business Studies (this corpus is also discussed in more detail in section 4.4.3 below). While these studies compare occurrences of the *forms* of different lexical bundles in different disciplines, they do not compare the *functions* of lexical bundles in different disciplines: a bundle is assumed to have the same function whichever discipline it occurs in, and regardless of its location in the text in which it occurs.

This section will compare the functional frameworks of Biber and Hyland and then discuss the two main issues which require further investigation, both of which relate to register and genre: the possibility of lexical bundles and discourse functions being missed as a result of the incomplete text samples in the register corpora used to develop these frameworks, and the lack of comparability of data between disciplines.

3.3.1 Functional frameworks for lexical bundles in academic discourse: Biber (2006) and Hyland (2008)

Lexical bundles were originally proposed as structural, rather than functional units, and were used in a similar way to other linguistic features to illustrate differences between registers. Biber et al. (1999) reveal differences in the forms of lexical bundles between the register of academic prose and the register of conversation. The principal findings in the 1999 study were that the structures of lexical bundles in each register are different: "most lexical bundles in conversation are building blocks for verbal and clausal structural units, while most lexical bundles in academic prose are building blocks for extended noun phrases or prepositional phrases" (1999: 992).

Several years after describing the formal features of lexical bundles, Biber, in collaboration with other researchers (Biber, Conrad, & Cortes 2004), developed a framework (from an initial taxonomy proposed by Cortes (2004)) to categorise their functions. The term *function* "refer[s] to the meanings and purposes of the language, functions that try to provide texture or organise the discourse according to situations or contexts" (Cortes 2004: 422). This definition explicitly connects the pragmatic meaning of lexical bundles, i.e. their discourse functions, to the context of situation. The study by Biber et al. (2004) determined the function of lexical bundles in different registers, and grouped together lexical bundles with similar functions into categories.

The focus of Biber's study was still on register rather than genre, however, as the goal of the functional framework was to "describe the extent to which each register uses lexical bundles for each of these functions" (Biber, Conrad, & Cortes 2004). The original framework proposed by Cortes has three core categories of function which relate to Halliday's metafunctional categories of register discussed in section 3.2.1, namely the ideational, textual, and interpersonal:

> **Referential bundles** perform an ideational function; they help writers structure their experience and determine their way of looking at things. Expressions in this category are time, place, or text markers, such as *at the beginning of*, *the end of the*, or *at the same time.*
> **Text organizers** are word combinations used to express textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse. Some of the functions performed by these expressions are contrast (e.g. *on the other hand*), inference (e.g. *as a result of*), or focus (e.g. *it is important to*)
> **Stance bundles and interactional bundles** perform interpersonal functions. Stance bundles express attitudes that frame some other proposition; expressions such as (e.g., *I don't know why*, *are more likely to*). Interactional bundles are conversational word combinations used to express politeness or to report, as in *thank you very much*, and *I said to him.*
>
> (Cortes 2004: 41)

The framework has subsequently evolved according to the data in which the bundles have been found. Cortes' framework was developed further in the more detailed study of academic register by Biber et al. (2004), which in turn was further revised by Biber (2006). Since this last study focused on academic discourse and did not make any comparisons with other registers, some bundles occurred more frequently or less frequently than in previous studies, and so the categories into which bundles were grouped either became more detailed and contained further subcategories, or were subsumed into superordinate categories. For example, some of the "interactional bundles" mentioned above by Cortes, such as *thank you very much* and *and I said to him*, occur often in the register of conversation, but seldom occur in the register of academic discourse. The category to which these lexical bundles belong is therefore less relevant to a study of academic discourse, since few bundles would occur to justify its existence as a separate category. As a result, while Biber's latest framework (see Table 3.1 below) still has the same three core categories, it subsumes "interactional bundles" into the "stance bundle" category.

The other study of lexical bundles in academic discourse is by Hyland (2008), in which he suggests an alternative functional framework for lexical bundles based on

his register corpus of academic prose (this is a different corpus from his 1998 journal article corpus). Although this framework was developed from that of Biber (Biber, Conrad, & Cortes 2004; Biber 2006) Hyland also found the interactional categories arising from conversation data to be un-necessary for classifying lexical bundles in academic discourse. Biber's data from "service encounters, institutional texts, and so on ... seems to have yielded far more personal, referential, and directive bundles than my more research-focused genres" (Hyland 2008: 13). The communicative purposes of the spoken and written texts included in Biber's conversational register subcorpus are so different from those in texts in academic discourse that many of Biber's categories and the bundles in them are not comparable: "use of the same sub-groups would invite unproductive comparisons" (ibid.).

Hyland's functional framework, therefore, does not use the same category headings as Biber. His categories are grouped under the headings of "research", "text" and "participants", and these categories in turn contain "sub-categories which specifically reflect the concerns of research writing" (Hyland 2008: 13). He places "stance" as a sub-ordinate category of interpersonal "participant-oriented" bundles which refer to either the writer or reader.

The two frameworks are presented side by side in Table 3.1 below, which aims to show the overlap between their superordinate categories. The lexical bundle functions in Biber's framework were originally identified in the register-based T2K-SWAL corpus, but he applies them here to the bundles found in a single genre, the textbook. The issue of the overlap between lexical bundle functions in a register and lexical bundle functions in a genre is discussed in the next section (3.3.2 below). The functions in Hyland's framework are those performed by lexical bundles identified in his register corpus.

Both frameworks have a roughly "interpersonal" category which is used to show a writer's attitude towards and stance on what is being written about, evaluating the propositions made, and including the audience. While "stance" is superordinate in Biber's study, it is a subcategory in Hyland's study. Both frameworks have a superordinate "textual" category, in which bundles are used to organise the discourse, although Hyland's category includes some functions, particularly those

Functions of common lexical bundles in academic textbooks (Biber 2006: 166-168)	Functions of the top 50 lexical bundles in RAs, dissertations and MA theses (Hyland 2008: 13-14)
Stance bundles express attitudes that frame some other proposition 1 Epistemic stance – impersonal *are more likely to, by the fact that* 2 Attitudinal/Modality stance a) Ability/effort: Impersonal *it is difficult to, it is possible to* b) Importance: Impersonal *it is important to, of the most important*	**Participant-oriented** 1 Stance features – convey the writer's attitudes and evaluations *are likely to be, it is possible that* 2 Engagement features – address readers directly *it should be noted that, as can be seen*
Discourse organizers express textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse *on the other hand, at the same time*	**Text-oriented** – concerned with the organisation of the text and its meaning as a message or argument: 1 Transition signals – establishing additive or contrastive links between elements *on the other hand, in contrast to the* 2 Resultative signals – mark inferential or causative relations between elements *as a result of, it was found that* 3 Structuring signals – text-reflexive markers which organise stretches of discourse or direct reader elsewhere in text *in the present study, as shown in figure* 4 Framing signals – situate arguments by specifying limiting conditions *in the case of, with respect to the*
Referential bundles help writers structure their experience and determine 1 Identification *is one of the, is known as the* 2 Specification of attributes a) Quantity/mathematical expression *of the number of, the magnitude of the* b) Predicative *is equal to the, is given by the* c) Tangible framing attributes *the size of the, in place of the* d) Intangible framing attributes *in the case of, as a result of* 3 Time/place/text/other reference a) Place or institution reference *the united states and, in the united states* b) General location reference or framing *in the same direction, the surface of the* c) Text deixis *as shown in figure, in this chapter we*	**Research-oriented** – help writers to structure their activities and experiences of the real world: 1 Location – indicating time/place *at the beginning of, at the same time* 2 Procedure *the use of the, the role of the* 3 Quantification *the magnitude of the, a wide range of* 4 Description *the structure of the, the size of the* 5 Topic – related to the field of research *in the Hong Kong, the currency board system*

Table 3.1: Comparison of the functional frameworks of Biber (2006) and Hyland (2008)

relating to "framing", which Biber puts in another category. Biber's "discourse organisers" is a single category, while Hyland's similar text-oriented category is sub-divided into transitional, resultative, structuring, and framing signals. The category with most difference between the two is the ideational metafunctional category.

There are a number of questions arising from Biber's and Hyland's work on the functions of lexical bundles which the isotextual, genre-based study of fixed collocational patterns in the present study may be able to clarify. Since Biber (2006) and Hyland (2008) have already proposed two alternate functional frameworks for fixed collocational patterns in academic discourse, it is not appropriate for this study to put forward a third, competing framework arising from the fixed collocational patterns in the corpus. It is instead more relevant to learn more about the bundles in Biber's and Hyland's functional frameworks by comparing their results with the frequent fixed collocational patterns found in the present study, since up to now few attempts have been made to compare or replicate their results. The next section outlines the areas which this study will aim to clarify.

3.3.2 Lexical bundles and the fixed collocational patterns in this study

The first issue which this study could clarify concerns the overlaps between register and genre. The corpus data in which Biber et al. (1999) identified lexical bundles was collected to represent a register, but the discourse functions later assigned to these bundles all work at the level of genre. As I have argued, it was logical for Biber to do this since genres are instantiations of registers, and registers include genres. However, the T2K-SWAL register corpus did not contain complete texts, only samples, and so not all areas of the discourse were included. I have also argued that genres are staged and goal-oriented: texts in a genre have beginnings, middles, and ends, and discourse functions are linked to these stages. The sampling procedure used by Biber meant that some discourse features in texts from a particular genre might have been excluded from the register corpus, and so Biber's results might have excluded some lexical bundles which customarily occur in texts in that genre. The only lexical bundles to be assigned discourse functions were those bundles which occurred in his register data, leaving the possibility that some of the "missing"

bundles – those that occur in the parts of the texts which were not sampled and which were therefore excluded from his corpus – might have different functions to the ones he identified. His functional categories might as a result comprehensively reflect the discourse functions in a register, but might not fully represent the discourse functions of the genres in that register. A study of lexical bundles in the genre-based research article corpus in this study could, therefore, reveal some of these "missing" bundles, and their functions could be further investigated.

The second aspect concerns both Biber's and Hyland's way of identifying a bundle's function from context, since it is not clear how much context they use to determine such discourse functions. There is some indication that Biber's and Hyland's functions were determined from concordance lines, which provide a limited amount of context (see section 4.2 below). Some categories seem to equate the function of a lexical bundle with its lexical content: the lexical bundle *the number of*, for example, is termed a referential "quantifier" (Biber's referential category 2a in Table 3.1 above) because it can be seen from the concordance line that it contains the word *number*. The category of "stance bundles of importance" (Biber's stance category 2b in Table 3.1 above) similarly includes bundles such as *of the most important*. It would difficult, using concordance lines alone, to assign a function to a lexical bundle which contains no lexical words, such as *in off the*. It may be that looking at the textual location of a bundle and the wider context in which it occurs, using the heuristic framework suggested in the previous chapter, may give a clearer idea of its function.

A related issue with the functions of lexical bundles is the assumption by both Biber and Hyland that all occurrences of a bundle have the same function wherever they occur in the register, regardless of genre, discipline, or where they are located in the text in which they occur. Only the frequency of occurrence of a bundle in each discipline is compared, and does not allow for the possibility that the function of a lexical bundle in the text in which it occurs might be related to its location. For example, a "quantifying" lexical bundle might occur more in the results sections of a research article, or a "stance bundle of importance" might occur more in introduction or discussions sections. In order to test such hypotheses, neither an isolexical register corpus nor a non-isotextual genre corpus would be suitable. Instead an isotextual study of a genre corpus of research articles, such as in the present study,

should reveal more about the relationship between location in the text of a fixed collocational pattern and its discourse function.

3.4 Summary

This chapter has attempted to situate the fixed collocational patterns in the present study in relation to academic discourse, particularly the notions of genre and register. In discussing previous work on lexical bundles in academic genres, it has also revealed issues which the present study of fixed collocations in isolexical and isotextual corpora may be able to clarify. The isotextual version of the corpus is likely to be able to reveal fixed collocational patterns, such as lexical bundles, which have specific discourse functions in particular areas of the discourse. Little can be said about a lexical bundle which is found more often in a subcorpus containing many short full texts than in a subcorpus containing an equal number of tokens but which contains a few long or fragmented texts. This study will investigate whether fixed collocational patterns can be found with similar forms to the lexical bundles identified by Biber and Hyland. It will then be able to investigate whether lexical bundle specified as having a particular function in the discourse by Biber and Hyland has a similar discourse function in the present study, and it will then investigate whether this function influences the location where the bundle occurs in a text.

It is obvious from the above discussion that the construction of the corpus in any empirical study of language is crucial. The next chapter accordingly discusses aspects of the field of Corpus Linguistics which will be employed in this study, and then discusses the benefits a distinction between isolexical and isotextual corpora may bring.

Chapter 4 Corpus Linguistics: a methodology and theory of language

4.1 Introduction

Studies of frequent fixed collocational patterns are necessarily empirical. It would be impossible to discover the most frequent fixed collocational patterns in research articles in different academic disciplines solely using intuition. It has already been pointed out in section 2.3.1 that language users are not always reliable at recalling the most frequent collocations of a word when asked to do so. It was argued that while some collocational patterns are sometimes easily retrievable from a researcher's intuition, for example *stark raving mad,* or *stark raving bonkers*, intuition is an unreliable basis for their systematic study.

Instances of collocational patterns are instead most comprehensively retrieved via computerised analysis of large amounts of language data. Indeed, it is "difficult and arguably pointless to study such things except through using large amounts of real data," (Moon 1997). It would be impossibly time-consuming to read all the data in order to find examples. Studies such as the present one are therefore computer-mediated, using large computer-held corpora to find and count attested occurrences of collocational patterns in use by many language users.

The construction of the corpus used in this study would not have been feasible ten years ago. Since the advent of the early computerised corpora of the 1960s, the amount and variety of the language data available electronically has grown exponentially, as has the amount of computer storage space and processing power available to individual researchers. Corpora produced as recently as the mid 1990s were limited by constraints which no longer pertain. The increase in capacity of the average home or office desktop PC in the last five to ten years has meant that large collections of texts can be built by individual researchers. Language data stored in computer memory can be retrieved, manipulated, and re-presented to researchers quickly.

This chapter first explains why I decided to construct a corpus for the present study, as none of the many large corpora which have already been constructed and which are available to researchers were suitable. It describes the composition and content of these existing corpora in some detail and shows how the incomplete or unequal numbers of texts and tokens in their different sections means that they cannot be used for comparative studies between academic disciplines. Having discussed these limitations, this chapter will then describe the design of the specialised research article corpus constructed as part of this study. The final section draws on work from Chapter 2 and Chapter 3 in comparing the research aims and methodology of this study with previous studies of word combinations in academic discourse.

4.2 The methodological approach

4.2.1 Introduction: corpus-based and corpus-driven approaches

Corpus linguists use large collections of textual data to investigate language, although there are different perspectives on what can be found in this way. Two distinct approaches to using corpora in linguistic research have been identified by Tognini-Bonelli (2001: 10): *corpus-based* and *corpus-driven*. A corpus-based approach is primarily a methodological application which uses a corpus as " a repository of examples to expound, test, or exemplify given theoretical statements" (ibid.). In contrast to corpus-based research, corpus-driven approaches use corpora to challenge existing views of language rather than confirm them. The corpus is used to generate, then test, then adjust or confirm hypotheses. Thus "a theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it" (Tognini-Bonelli 2001: 11). This approach is more than purely methodological, as it has the potential to offer insights into language which have theoretical implications. This section outlines some of the drawbacks to some of the features of a corpus-based approach, and explains why, in order to identify the fixed collocations to be investigated, this study will take a corpus-driven approach.

4.2.2 The limitations of annotated data

One methodological feature of corpus-based work is that data is usually annotated with tags: labels within brackets which corpus software uses but does not process as being part of the actual corpus text. These tags contain extra information, such as the part of speech for each word or the constituent role of each word in a sentence after parsing. Other information can indicate anaphoric reference or semantic field (Hunston 2002: 87-88). While annotation has a number of advantages and disadvantages, the corpus in the present study will not be annotated. Annotation is useful for learning more about features such as parts of speech, for which there are existing analytical frameworks and for which tagsets have been developed. More can be learned from a corpus about the finite base verb form of *break,* for example, if the corpus data has been annotated with part-of-speech tags. This allows the researcher to look at examples of *break* which are used as finite verbs, i.e. *<VVB>break,* without having to sort and discard instances of *break* which are used as infinitives (*<VVI>break*) or singular nouns (*<NN1>break*).

One drawback to using a pre-annotated corpus is that the researcher views the data through the prism of existing theories. If a corpus is tagged according to word classes, for example, then the theory of word classes is not likely to be changed by analyses of that corpus which relies on these tags. As pointed out by Sinclair, many such tagging systems used for corpora are "based on problematic analytical systems derived from traditions of language analysis which have, in the past, rejected corpus evidence" (1991: 22). Annotated data limits "not the kind of question that can be asked, but the kind of question that usually is asked" (Hunston 2002: 93).

In a corpus-driven approach, using non-annotated corpus data means that existing categories are not taken as given, and that researchers can notice "the strength of patterning which emerges from the rawest of unprocessed data" (Sinclair 1991: 117). This can be illustrated by the 'new' word classes identified by Hunston and Francis (2000: 181). These are mostly sub-classes of the existing classes of nouns and verbs, and would not necessarily have been revealed if the corpus data was tagged and searched according to existing word classes. "Variable nouns", for example, are nouns which have both countable and uncountable forms, such as *injustice* or *lunch*,

but whose patterns of behaviour are observed to be sufficiently different from that of other countable and uncountable nouns that Hunston and Francis class them separately. The corpus-driven work of Hunston and Francis has led to a reformulation of the theory of word classes.

Another form of annotation used in corpus-based work is lemmatisation, in which different word forms are given the same tag because they are seen as belonging to the same lemma. In the case of *was* and *were* the lemma is *BE*, and so in the BNC all instances of *was* and *were* are tagged as the past tense forms of the verb BE, i.e. <*VBD*>. Oakes (1998: 153) recommends "pre-preparing" a corpus in this way so as to prevent orthographically distinct instances of the same word such as *be* and *was* being listed separately. To Sinclair, on the other hand, listing occurrences of such word forms together as instances of the same lemma risks obscuring the evidence of how individual word forms behave. "Each distinct form of a verb is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of flexibility" (1991: 8). Corpus-driven approaches do not therefore lemmatise data.

An example of how annotation determines what the corpus can or cannot easily reveal is the case of *it*. In the BNC all occurrences of *it* are tagged as <w PNP>, i.e. as a pronoun with reference to another noun in the text. However, those occurrences of *it* which function very differently, namely as a "dummy subject," are also annotated with the same <w PNP> tag, as in Figure 4.1:

Though the Report could be interpreted as advocating an exchange rate union, it is now accepted that it recommends a currency union.
<s n=112> <w CJS>Though <w AT0>the <w NN1>Report <w VM0>could <w VBI>be <w VVN>interpreted <w CJS-PRP>as <w VVG>advocating <w AT0>an <w NN1>exchange <w NN1>rate <w NN0>union<c PUN>, <w PNP>it <w VBZ>is <w AV0>now <w VVD-VVN>accepted <w CJT>that <w PNP>it <w VVZ>recommends <w AT0>a <w NN1>currency <w NN0>union<c PUN>.

Figure 4.1: A sentence from the BNC with and without tags

The tagging of *it* in this way means a researcher cannot use tags to distinguish between anaphoric or catephoric *it* and introductory or dummy *it*. The different uses of *it* would have to be separated manually.

The field of learner corpora also illustrates the problems arising from a corpus-based approach using annotated data. Errors made by learner writers are tagged according to an error tagset so that more can be learned about the types of errors made by particular learners. A problem arises due to the fact that erroneous writing can only be identified in terms of what would be seen as correct, and, since the learner is not present, there needs to be an element of guesswork about what the learner was unsuccessfully trying to mean. It is difficult to decide when something that has been identified as "an error" has been corrected sufficiently for it to become "not wrong": there is obviously no such construct as "a correct". Dagneaux et al. (1998: 165), for example, admit that "... the quality of error correction ... remains problematic because there is regularly more than one correct form to choose from. The inserted correct form should therefore rather be viewed as one possible correct form - ideally the most plausible one than as the one and only possible form."

A specific example of the limitations of error tagging are illustrated in Figure 4.2 below, which shows a concordance of the occurrence of a particular class of lexico-grammatical error, XNPR, in Dagneaux et al. (1998: 168):

complemented by other	(XNPR)	approaches of $approaches to$ the subject. The written
are concerned. Yet, the	(XNPR)	aspiration to $aspiration for$ a more equitable society
can walk without paying	(XNPR)	attention of $attention to$ the (LSF) circulation $traffic$.
could not decently take	(XNPR)	care for $care of$ a whole family with two half salaries
be ignored is the real	(XNPR)	dependence towards $dependence on$ television.
are trying to affirm their	(XNPR)	desire of $desire for$ recognition in our society.
such as (GA) the a	(XNPR)	drop of $drop in$ meat prices. But what are these
decisions by their	(XNPR)	interest for $interest in$ politics. As a conclusion we can
hope to unearth the	(XNPR)	keys of $keys to$ our personality. But (GVT) do scientists
and (GVN) puts put	(XNPR)	limits to $limits on$ the introduction of technology in their
This dream, or rather	(XNPR)	obsession of $obsession for$ power of some leaders can

Figure 4.2: Output of search for XNPR (Dagneaux et al. 1998: 168)

The final concordance line in Figure 4.2 shows that *obsession of* has been diagnosed as a lexico-grammatical error, and has been corrected to *obsession for*.

```
This dream, or rather (XNPR) obsession of $obsession for$ power of
some leaders can
```

Although *obsession for* may be a possible correction at this point, it may not be the most appropriate correction. In the BoE, 2,770 of the 5,241 occurrences of *obsession* are followed by *with,* in contrast to the 364 occurrences followed by *for*, (the 131 occurrences of *obsession of* notwithstanding) so that *obsession with power* is much more common. However, both *obsession for power (2)*, or *obsession with power (4),* both of which are possible combinations, are much less frequent than *lust for power* (72 occurrences) or *appetite for power* (21) or *desire for power* (28). The learner's error may not have been with the use of *obsession* at all; rather it was a lack of knowledge of the backward collocates of *power.*

The focus of the present study of fixed collocational patterns means that it is necessarily corpus-driven, and so the fixed collocational patterns to be investigated will emerge from the corpus without any prior annotation or manipulation of the data. According to Sinclair's open choice/idiom principle discussed in section 2.3.1, any point in a text is either part of a collocational pattern or not, and so the point in the text which has been tagged may be part of a collocational relationship that extends backwards as well as forwards. These relationships will be missed by manual or automatic tagging which only works on text forwards from left to right. Instead of identifying tags, the corpus software can be instructed to identify strings of characters and spaces of set lengths occurring above a particular frequency. There is also little point in tagging frequently occurring fixed collocational patterns as items themselves, since they can only be tagged once they have been identified as frequently occurring fixed collocational patterns. Once they have been identified in the present study, then there is no further need to tag them.

4.2.3 Viewing context: the limitations of concordance lines

Another methodology traditionally used in corpus-driven approaches, but of limited use for this study, is the manipulation of corpus data by computer concordancing software. The software is instructed to search for a given string of characters in the electronic text files comprising the corpus, and then presents the results in the form of concordance lines. A concordance line is a line of around 70 characters which

shows the search item in the centre of the screen. Most screens are wide enough to allow between 5 and 7 words of context to be displayed either side of the search item. This limit is arbitrary, arising from limits of the technology used in the 1960s and 1970s: this was as much information as a computer screen could display and which a printer could print (Johns, King, Leech, personal communication).

Concordance lines are essential for showing the researcher the linguistic patterns entered into by individual words in context. Figure 4.3 below shows two miniaturised displays of a set of concordance lines for *stark* (Partington 1998: 9) in which the lines are arranged in alphabetical order. Words immediately to the right of *stark* are sorted so that words nearer the beginning of the alphabet are higher up the screen. On the left of Figure 4.3, the data after this manipulation is presented in Courier New, a non-proportional font in which every character takes up the same space. On the right it is in Times New Roman, a proportional font which does not reveal the regularity around stark and its environment (i.e. the fixed collocational pattern *in stark contrast to*) in the same way.

The choice of font is crucial to whether the information about the data presented in the concordance lines can be noticed by the researcher. In order to identify linguistic features in concordance lines the researcher does not use any cognitive linguistic processes. He or she does not engage linguistically with the data until his or her brain, using cognitive procedures not connected to language recognition, has noticed the graphical regularity (circled). "The computer is valuable in stressing the difference between the text as a physical object and the way it is perceived by a language user" (Sinclair 1991: 7).

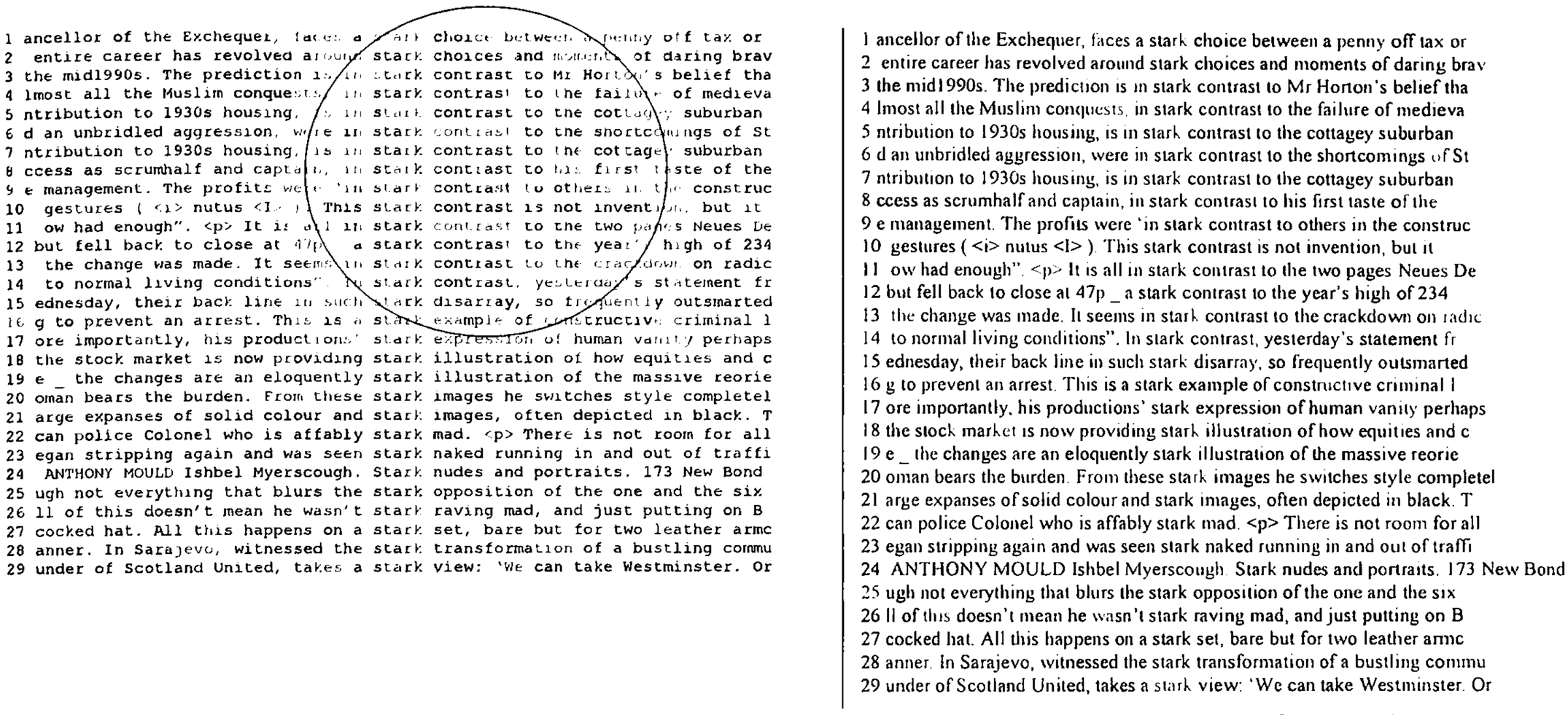


Figure 4.3: Concordance of *stark* (Partington 1998: 9) in Courier New (left) and in Times New Roman (right)

Concordance lines can thus reveal the "strict interconnection between an item and its environment" (Tognini-Bonelli 2001: 11), both syntactic and paradigmatic, something which not all empirical studies of language use in academic discourse take into account. Jasienski (2006), for example, in a test of whether natural scientists admit surprise at their findings more frequently than social scientists or non-scientists, searched for "words indicating surprise" among 30 million abstracts of English-language scientific papers and 8 million articles in the Social Sciences and Arts & Humanities citation indices. He compared the frequencies of words appearing in these abstracts with their frequencies in the Brown corpus and found that "the word 'surprising' appears 12 times more frequently in the natural sciences than in standard English and 1.3 times more frequently than in social sciences, arts and humanities," and concludes from this observation that:

> the study of nature does indeed seem to surprise us. The odds of finding in abstracts of scientific research papers a result or conclusion described as 'surprising', 'unexpected', or 'unusual' are an order of magnitude greater than in standard language and several times greater than in non-science academic abstracts.
> (Jasienski 2006: 1112)

Jasienski's conclusion is not necessarily supported by the evidence, however. If one uses a corpus to search for *surprising* and find *n* occurrences, then, before one makes any claims about what this finding means, one would do well to check its textual environment, namely whether *not* is used somewhere in the span to the left of

surprising. All the examples below (from natural science journals) use the word *surprising*, and which thus would be counted in Jasienski's results, but only in Example 4.1 and Example 4.3 is surprise actually expressed in the text:

> Example 4.1: Given that Hrp1p is required for nuclear mRNA polyadenylation, it is surprising that the cell survives Hrp1p relocalization to the cytoplasm during hypertonic conditions. (0070040140934 Cell Biology)

> Example 4.2: The experimentally determined losses have relatively large error bars at the highest T, so it would not be surprising if the nominal result for int (300 K) is somewhat too high. (0030030940310 Electrical Engineering)

> Example 4.3: These two characteristics are quite surprising actually and therefore they should be added to the increasing collection of phantom weird properties. (0090056970112 Nuclear Physics)

> Example 4.4: Considering gravitational effects, the larger is d, the more efficient are gravity effects to deform the interface. It is thus not surprising that inertia l effects have to be increased to compensate for it. (0060011230328 Mechanical Engineering)

Jasienski's generalisation from his results is therefore unfounded. *Surprising* may indeed be used much more often in Natural Science abstracts than Social Science abstracts, and even more often than in the Brown corpus, as he claims, but it does not necessarily follow from this observation that natural scientists "admit surprise at their findings" any more than do social scientists or writers in standard English.

This study will use concordancing software to investigate the immediate co-textual environment of fixed collocational patterns, such as the noun groups they appear to frame, but the limited width of concordance lines means they are less useful for examining the wider context of a pattern. When the search item is a string of words, more of the screen is taken up with the search item and there is less room for the co-text. Academic discourse has a well known tendency to contain long complex noun groups (Biber 1988: 268; Cohen et al. 1988; Carter & McCarthy 2006). Such nominalization is used frequently because it allows processes and developments which are commonly expressed in clauses to be treated as concepts. Switching from a clause, such as in Example 4.5, to a nominalised form, such as Example 4.6, makes it easier to evaluate as part of the developing argument in the discourse (Kress 1989).

> Example 4.5: The incidence of tuberculosis in England and Wales suddenly increased between 1988 and 1992.

> Example 4.6: Bhatti et al., for example, discuss the sudden increase of the incidence of tuberculosis in England and Wales between 1988 and 1992. According to these authors the recent entry of African refugees was not of major importance for the national tuberculosis increase.

The length of such noun groups in academic discourse means that any lexical or syntactic patterns in their surrounding co-text are harder to spot visually since they are off-screen, a point illustrated by Figure 4.4 below. This shows concordance citations for *evaluations* in an Economics research article sorted alphabetically up to three words to the left. It can be seen that repeated regularities continue up to four places to the left of the sorted words because of the frequent occurrence of the eight-word noun group *the relationship/correlation between tip sizes and service evaluations*. The regular patterns displayed on screen are largely those of the constituents of this noun group and do not allow other clause elements to be fully seen.

```
14 service evaluations after statistically controlling for food evaluations; 4) the differen
15 s calculated and used in tests of the effects involving food evaluations. For studies wit
16 ervice evaluations and the correlation of tip size with food evaluations; 5) the zero-ord
17  was not significantly related to servers' or third-parties' evaluations of the service (
18 ; 6) the correlation between patronage frequency and service evaluations; 7) the partial
19 and the relationship between patronage frequency and service evaluations was significantl
20 8) the interaction effect of patronage frequency and service evaluations on tip size. Tes
21 s of the correlation between patronage frequency and service evaluations. Meta-analyses o
22 though the average relationship between tip size and service evaluations was statisticall
23  within-restaurant relationship between tip size and service evaluations more than on the
24 ically significant relationship between tip size and service evaluations. However, that r
25 nges. However, the relationship between tip size and service evaluations was weaker than
26 xplanation for the correlation between tip sizes and service evaluations is that tip size
27 vailable, partial-correlations between tip sizes and service evaluations that statistical
28  not moderate the relationship between tip sizes and service evaluations and that server
29 s should be comparable to that between tip sizes and service evaluations. Tests of the pr
30 ificant positive relationships between tip sizes and service evaluations (Lynn and Grassm
31 ns; 7) the partial correlation between tip sizes and service evaluations after statistica
32 s conclusion, the relationship between tip sizes and service evaluations remained signifi
33 od; 3) the partial correlation between tip sizes and service evaluations after statistica
34 explanations for the correlation between tipping and service evaluations. These findings
35 effects of customer' patronage frequency on tips and service evaluations rather than to a
36 ct of some third variable's effects on both tips and service evaluations. One likely cand
37 es (percent tips or bill-adjusted residual tips) and service evaluations. Tests of this r
38 mall, but reliable and positive relationship between service evaluations and tip sizes an
39 alculated and used in tests of the effects involving service evaluations. Similarly, in c
40 ionship is due to the effects of customers' moods on service evaluations and tips rather
41 .46, one-tailed p < .0004, and marginally related to service evaluations, r = 0.03, z = 1
42 ying the service-tipping relationship is specific to service evaluations and that custome
```

Figure 4.4: Concordance of *evaluations* in an Economics research article, sorted 3 words to the left (Oakey 2005: 176)

Having discussed the benefits of a corpus-driven approach in relation to a corpus-based approach, this chapter moves on to discussing the written academic English corpus data already available, and why it was not seen as suitable for the present study.

4.3 Limitations of existing corpora of written academic English

There are two possible methodological approaches to corpus-driven comparisons of the frequency and role of fixed collocational patterns: either data in existing representative "off-the-shelf" corpora can be grouped into categories reflecting different areas of language use, or original corpora of data chosen from different categories can be constructed. As discussed in Chapter 3, various overlapping labels can be used for these different categories: comparing different *registers* (Biber et al. 1998: 135) reveals linguistic features typical of different situations of use, such as purpose, topic, and setting; making comparisons between texts belonging to different *genres* (Lee 2001: 46) can show how language varies according to socio-culturally determined factors.

However, while there are several sources of academic writing data available in existing off-the-shelf corpora, the design principles underlying these corpora mean that their suitability for both register and genre-based comparative work on collocational patterns is limited by two considerations. First, while it is possible for such corpora to be sub-divided to allow the study of language features across different registers or genres, the amount of data in each subcorpus is often unequal. Second, some corpus compilers achieve a balanced number of tokens by using text fragments, with the result that discourse features from the parts of the texts which are not included in the sample are not represented in the corpus, which in turn risks biasing the results of searches of the corpus. This section discusses these limitations in more detail, and then shows how they might be addressed by introducing a distinction between isolexical and isotextual versions of a corpus. As I have already suggested, one version will allow results to be compared across a similar number of tokens, while the other will allow results to be compared across a similar number of texts.

4.3.1 The Brown corpus tradition: Brown, LOB, FLOB, FROWN and BNC

The corpora in this tradition date back to the first attempt to make a representative computerised corpus, the Standard Corpus of Present Day American English compiled by Francis and Kučera at Brown University in 1964 (Francis & Kučera 1964/1979). This corpus, known as the "Brown" corpus, was intended to be a representative sample of American written English, what Sinclair (2005: 3) calls a *normative* corpus. Statements made about linguistic features in the corpus were to be accepted as reflecting American written English as a whole. The data collected for the Brown corpus had to satisfy a number of criteria in order to be included. First, most obviously, the texts had to be written in English by an American author. Francis and Kučera were confident that "so far as it has been possible to determine, the writers were native speakers of American English" (1964/1979: 1). The second criterion concerned the subject matter of the texts in the corpus; the compilers had to identify the topics which were written about in English by Americans, and then decide how much of each subject to include in the corpus so that it represented American written English.

The process of attempting to represent all of a particular language in this way is in danger of becoming circular. In order to accurately replicate a language in microcosm in a corpus, the compilers face two problems: which varieties to include, and how much of each variety to include. They need to know how many different varieties of language - such as mode (spoken or written), utterer (male or female), register (formal or informal) - are used in the macrocosm, i.e. the language as a whole: their corpus must include samples of each variety. Second, they need to know the proportion of the language accounted for by each variety, in order to ensure that the proportions of the different types of language in the microcosm reflected the proportions in the macrocosm. Since most languages are spoken and written by many millions of people, these proportions are almost impossible to determine for a whole language. To be able to determine these proportions accurately, researchers would have to investigate the language as a whole, and in order to do this they would need a representative corpus, and so the process begins again from the same point.

Rather than attempting to represent the whole English language, Francis and Kučera aimed to represent American written English as a "standard" against which to compare other instances of language use. They were thus able to solve the first problem by consulting Brown University's librarians to decide on a classificatory system that would include every variety of American written English. This approach yielded two basic classes of American written English texts: "informative" and "imaginative" prose, which were in turn subdivided into smaller categories. "Informative" prose consisted of Press (Reportage, Editorial, and Reviews), Religion, Skills and Hobbies, Popular Lore, Belles Lettres (Biography, Memoirs, etc.), Miscellaneous, and Learned. "Imaginative" prose consisted of General, Mystery and Detective, Science Fiction, Adventure and Western, Romance and Love Story, and Humor. These are library-derived subject categories; they reflect a concern with the topic of a text, rather than with context-of-situation factors such as the form of a text or the intended audience it was written for, and so the Brown corpus categories reflect subjects rather than genres or registers.

Francis and Kučera tackled the second problem, of making sure that the quantities in the corpus reflected the quantities actually in the language as a whole, by canvassing the opinions of distinguished linguists[1] as to how many texts from each category should be included. They averaged out the various answers and then collected 500 random samples of 2000 words from texts, restricted to those written in 1961, in each of these categories (Francis & Kučera 1964/1979).

The LOB corpus, a collocation of one million words of British English written in 1961 continued the Brown tradition of corpus design and construction in the 1970s. This aimed to mirror the Brown corpus as closely as possible in order to allow comparisons between British and American English (Johansson et al. 1978). Categories and sampling criteria for the two corpora were largely unchanged; Figure 4.5 below illustrates the composition of the two corpora. Linguists at the University of Freiburg updated these two corpora with texts written in the early 1990s to create the FROWN and FLOB corpora (Hundt et al. 1998; Hundt et al. 1999). The largest corpus constructed in the Brown tradition, the British National Corpus (BNC) (1995;

---

[1] John B. Carroll, W. Nelson Francis, Philip B. Gove, Henry Kucera, Patricia O'Connor, and Randolph Quirk

revised as the *BNC World* 2000), contained 90 million written and 10 million spoken words of British English from 1991. It contained texts in similar categories to the Brown/LOB corpora, but which were termed *subject domains*.

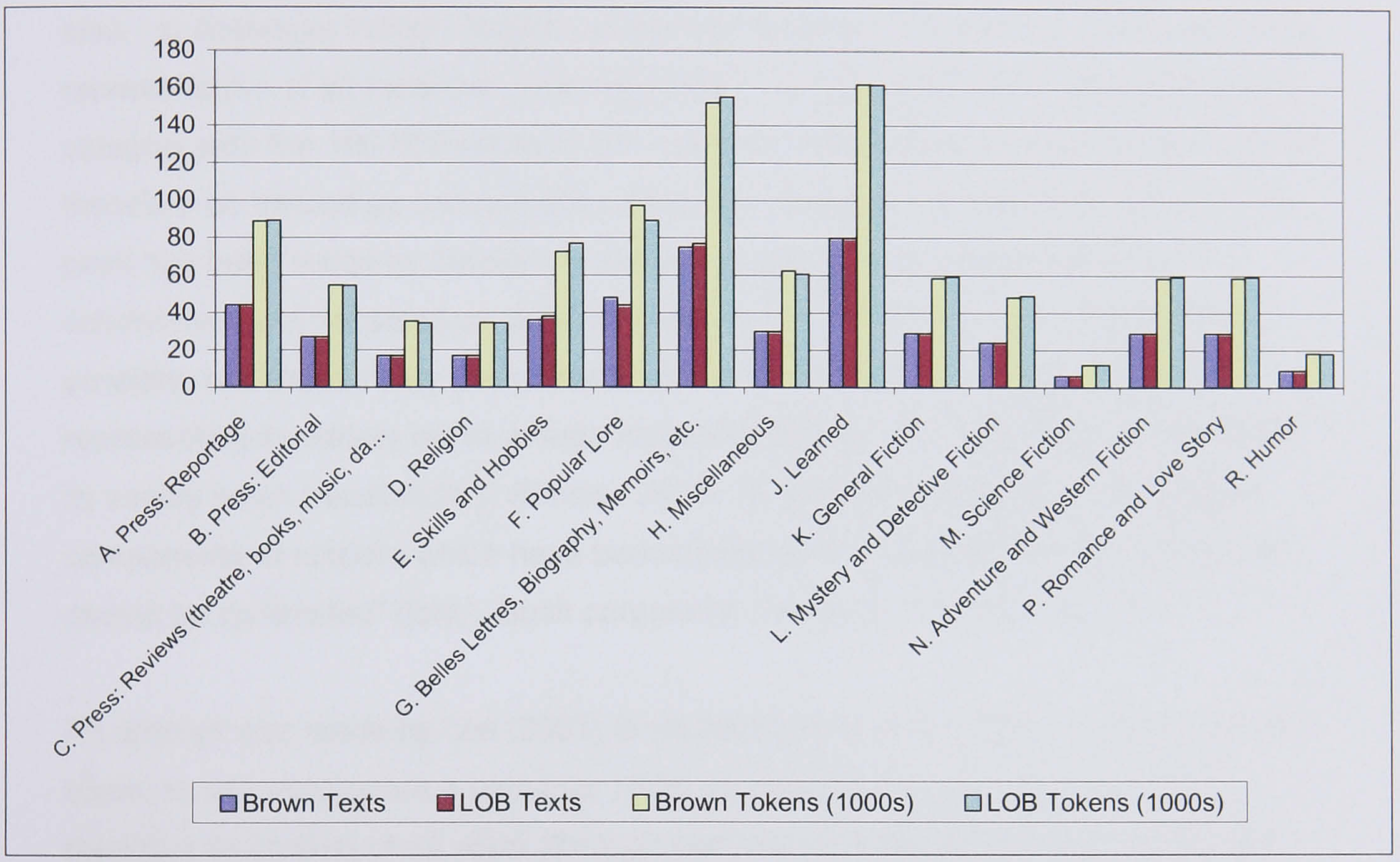


Figure 4.5: Distribution of text samples in the Brown Corpus and the LOB Corpus

The Brown, LOB, and BNC are very small in relation to 21st century normative corpora, such as those of commercial publishing houses such as the Collins Word Web and the Cambridge International Corpus, which contain hundreds of millions of tokens; and they can only be said to be representative of the written American and British English of one or two generations ago. Despite this, the systematicity with which they were constructed and annotated, as well as the volume of scholarly work based on them, has meant that they are still widely used as reference corpora and for the training of automatic text parsing software.

Apart from the advancing age of their data, however, their main drawback for the purposes of this study is the status of the language data in their respective categories or domains. If language in different categories is collected as a normative

sample of a language, none of the individual categories can on their own be representative of that category in the language as a whole, simply because they were collected as part of a larger representative whole sample. The Brown corpus "Learned" category of 160,000 words, for example, consists of samples from academic texts which were collected in order to go towards representing something else, i.e. American Written English as a whole. It cannot be seen on its own as being representative of all American Learned writing. Comparisons of the language of this category with the 160,000 words in the "Learned" category of the LOB corpus cannot therefore be treated as comparisons of the two categories of "Learned" writing. This point has been made by Sinclair: "a sample corpus will be discontinuous; its subdivisions will not stand as samples themselves," (1991: 24) and "it is perfectly possible, and indeed very likely, that a corpus component can be adequate for representing its variety within a large normative corpus, but inadequate to represent its variety when freestanding" (Sinclair 2005: 3). Sinclair stressed that "only those components of corpora which have been designed to be independently contrastive should be contrasted" (ibid.). Such corpora he termed *contrastive* corpora.

An attempt was made by Lee (2001) to reclassify texts in the BNC in order to make it easier to see what genre a particular instance of language use came from, and therefore its context of situation (through features such as author type, author age, author sex, audience type), and to allow researchers to isolate subcorpora from the BNC in order to compare different genres. Lee proposed a reclassification of the BNC text samples from its domains (Imaginative, Natural Science, Applied Science, Social Science, World Affairs, Commerce, Arts, Belief & Thought, Leisure, and Unclassified) into 'genres' (such as academic and non-academic, administrative, advertising, biography, broadsheet news).

While Lee's genre categories were seen as helpful, and were incorporated into the 2000 World edition of the BNC, they also were based on an underlying assumption that a normative corpus could be divided into a contrastive one. It was not the intention of the BNC designers, however, for it to be used for contrastive studies between or within different genres (Aston 2001). The number of texts, and consequently the number of tokens, in each genre category is unequal. The study in Oakey (2002a) used the "academic writing" genre of the BNC based on Lee's

categories to compare occurrences of lexical phrases in three different academic disciplines within this genre. However, the wide disparity in the number of texts and numbers of words in the three disciplines, as can be seen in Figure 4.6 below, meant that the results needed to be normalised for comparisons to be made.

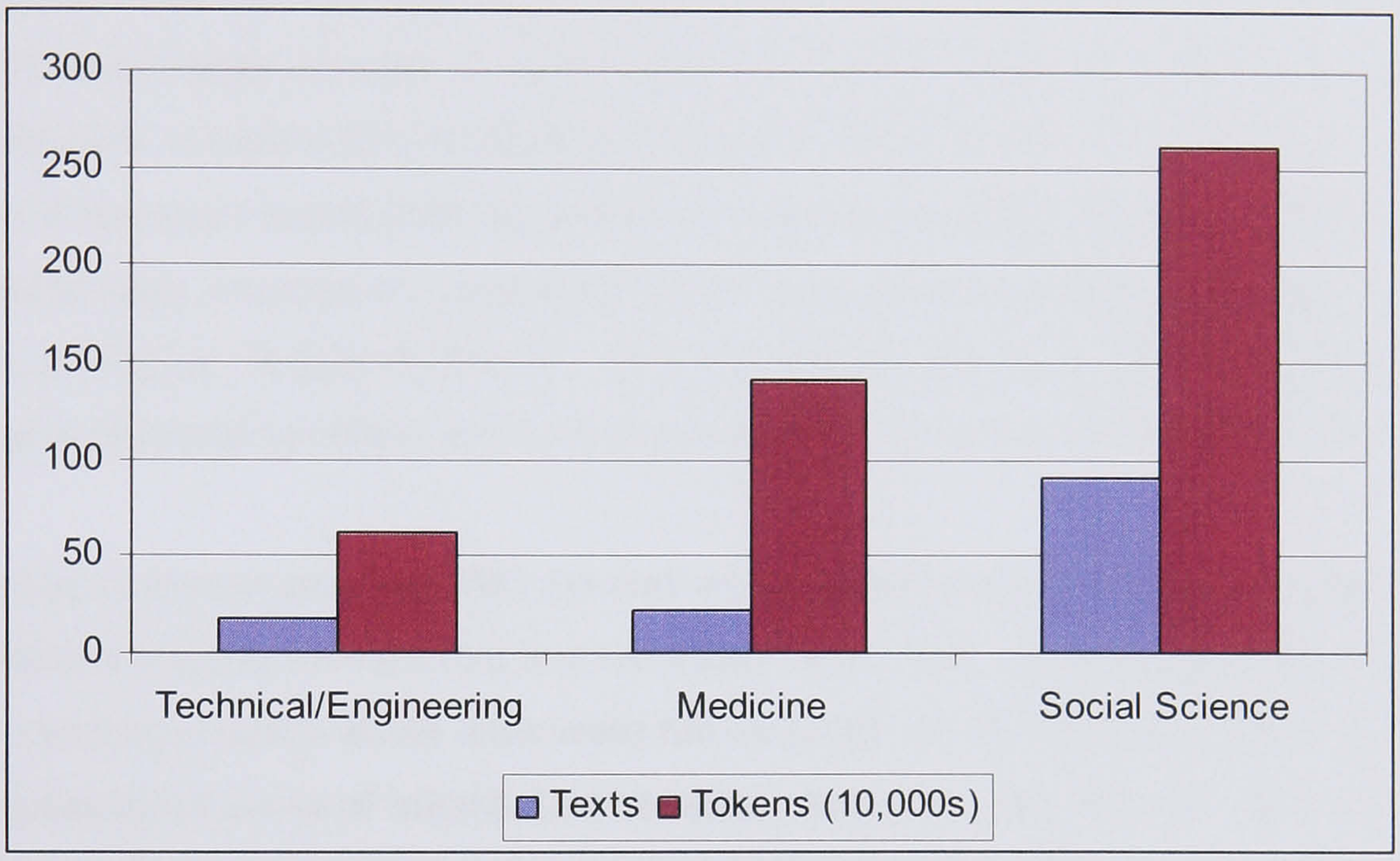


Figure 4.6: comparison of three different academic writing sub-corpora in the BNC

For occurrences of lexical phrases to be compared between these three disciplines, results either had to be normalised down to the level of the subcategory with the fewest tokens, in this case the approximately 500,000 tokens of Technical and Engineering writing, or normalised up to the level of the sub-category with the most tokens, namely the 2,500,000 or so tokens of Social Science writing. The former method involved dividing the results for Medicine roughly by three and those for Social Science roughly by five. The latter method would have involved multiplying results for Technical and Engineering writing by five, and those for Medicine by roughly three. In this latter case, the results would have been highly dubious, because the "long tail" distribution of language data (see section 4.3.3 below) means that the most frequent words in a text occur very frequently in relation to the infrequent words in the text. It cannot therefore be assumed, for example, that if a lexical phrase occurs *n* times in 500,000 tokens of Technical and Engineering writing, it will occur 5*n* times in 2,500,000 tokens. Similarly it cannot be assumed that

if a lexical phrase occurs *n* times in 5,000,000 tokens of Sociology writing, it will occur *n*/10 times in 500,000 tokens. It is thus essential for the corpus for the present study to contain a comparable amount of data so that like can be compared with like.

When written data from a single text type, the periodical, is isolated from the BNC using Lee's criteria, the resulting subcorpora are even more unbalanced. The proportionally larger amount of words taken from two journals, *Gut* and the *Weekly Law Reports*, is particularly noticeable, as can be seen on the right of Figure 4.7 below. It has been found that this imbalance causes normally infrequent words such as *gastric* (adj), *mucosa* (n), and *colitis* (n) to "leap up into the top 8000 frequent words of English" (Kilgarriff 2002: 1). This also affects research into the discourse functions of lexical bundles, and so this point will be returned to in section 4.4.3

The second reason why the BNC (as well as the Brown and LOB) is unsuitable for the present study is the fact that text samples, rather than whole texts, were used. In many existing corpora whole texts were not included, as in order to keep the sample size balanced in terms of tokens, text samples rather than whole texts were used. The inclusion of text fragments in a corpus means that the data is in some way incomplete (Sinclair 1991: 23-24; L. Flowerdew 2004: 15). Meaning as context of situation by definition cannot be investigated in such a corpus because not all of the situation, namely the text, is included. The Brown tradition's sampling process leads to fragmented texts; this invalidates their status as texts since they are no longer whole. If a text is to be studied as a communicative act in a particular context of situation, then all language used to perform that act needs to be available for study.

As a brief illustration of this difficulty in relation to the present study, we can propose the existence of a collocational pattern which plays a role in the introduction section of texts. If a corpus consisting of randomly-selected text fragments is searched in order to test this hypothesis, the results will inevitably prove inconclusive, since the sampling construction method may or may not have included the introduction sections of the texts in the corpus. Since the present study is not only quantitative (counting and comparing frequencies of fixed collocational patterns) but also qualitative (attempting to describe the use of these patterns), it needs to study full texts which include all the linguistic aspects of each communicative act.

4.3.2 The MicroConcord B corpus

The next available corpus of written academic English, the *MicroConcord B* corpus, is a collection of "more-or-less academic texts" (Murison-Bowie 1993: 53) supplied with *MicroConcord* software (Scott & Johns 1993) as an attempt to make corpora accessible to language teachers and learners (Johns 1986; Thurstun & Candlin 1997; Partington 1998). While the corpus contains data from five different disciplines (see the left of Figure 4.7 below) allowing comparison of the language used in different subjects, the individual sections are between 180,000 to 220,000 tokens each, which does not permit large-scale analysis of frequently occurring word combinations. Moreover, the texts in each area are fragments, leading to the problem mentioned in the previous section.

4.3.3 The Bank of English

Written academic English data is also available in the Bank of English. This corpus was the result of the COBUILD project, (Sinclair 1987) the first large-scale commercial corpus compilation project. COBUILD aimed to produce corpus-informed dictionaries and other English language teaching materials, and so its corpus is different in composition from corpora constructed along the lines of the Brown corpus. First of all, its designers eschewed sampling, and instead proceeded according to the principle of "more is better." This insistence on a large corpus was due to the fact that word frequencies are distributed unevenly, so that at least half the words in any text occur only once. This type of distribution was identified as a property of word frequency lists by Zipf (1935) in the pre-computer era, but has since been found - and named the "long tail" distribution - in other areas such as the sales figures of online retailers:

> I used to work at Amazon.com ... here's how I described the long tail to a marketing guy who didn't get it (and thought we should focus more on bestsellers): 'We sold more books today that didn't sell at all yesterday than we sold today of all the books that did sell yesterday.'
>
> (Brier 2005)

The long tail distribution of online book sales reflects the fact that online retailers such as Amazon.com make more money selling single copies of two million different

books a year than they would if they sold one million copies of a bestselling *Harry Potter* book annually.

The long tail distribution of words meant for Sinclair that a corpus had to be large enough to contain sufficient occurrences of the different senses and uses of infrequent words on which to base a dictionary entry. The definition of "large enough" has continued to expand. Moon (2007: 166) for example, points out that the entry for *skate* in the first COBUILD dictionary was based on 35 occurrences of the lemma, as opposed to the approximately 5,000 occurrences in the 450-million word 2001 BoE version.

While the "academic writing" portion of the BoE, which consists of around seven million tokens from American academic textbooks from several disciplines, is larger than those of the corpora in the Brown/BNC tradition, its principal limitation for the purposes of the present study is that there is little balance, both in terms of texts and tokens, between the disciplines represented (see Figure 4.7 below); comparisons between subject disciplines in this study would be hard to make. The disciplines represented are also biased towards non-Science disciplines.

Figure 4.7 below shows a graphical representation of the corpora discussed in this section and shows the number of texts and tokens in relation to each other: It can be seen that using academic writing from normative corpora would not be appropriate for the present study. A corpus consisting of comparable subcorpora was instead constructed, as will be discussed in section 5.2. The additional benefit to this study of a purpose-built corpus, in addition to the ability to control the number of texts and tokens in each discipline subcorpus, would be that I have the advantage of knowing what is in the data. Both Hoey (1997: 2ff.) and Lee (2001: 53) point out that mis-labelling or misclassification of files in off-the-shelf corpora can lead to compromised research results for other users of the data.

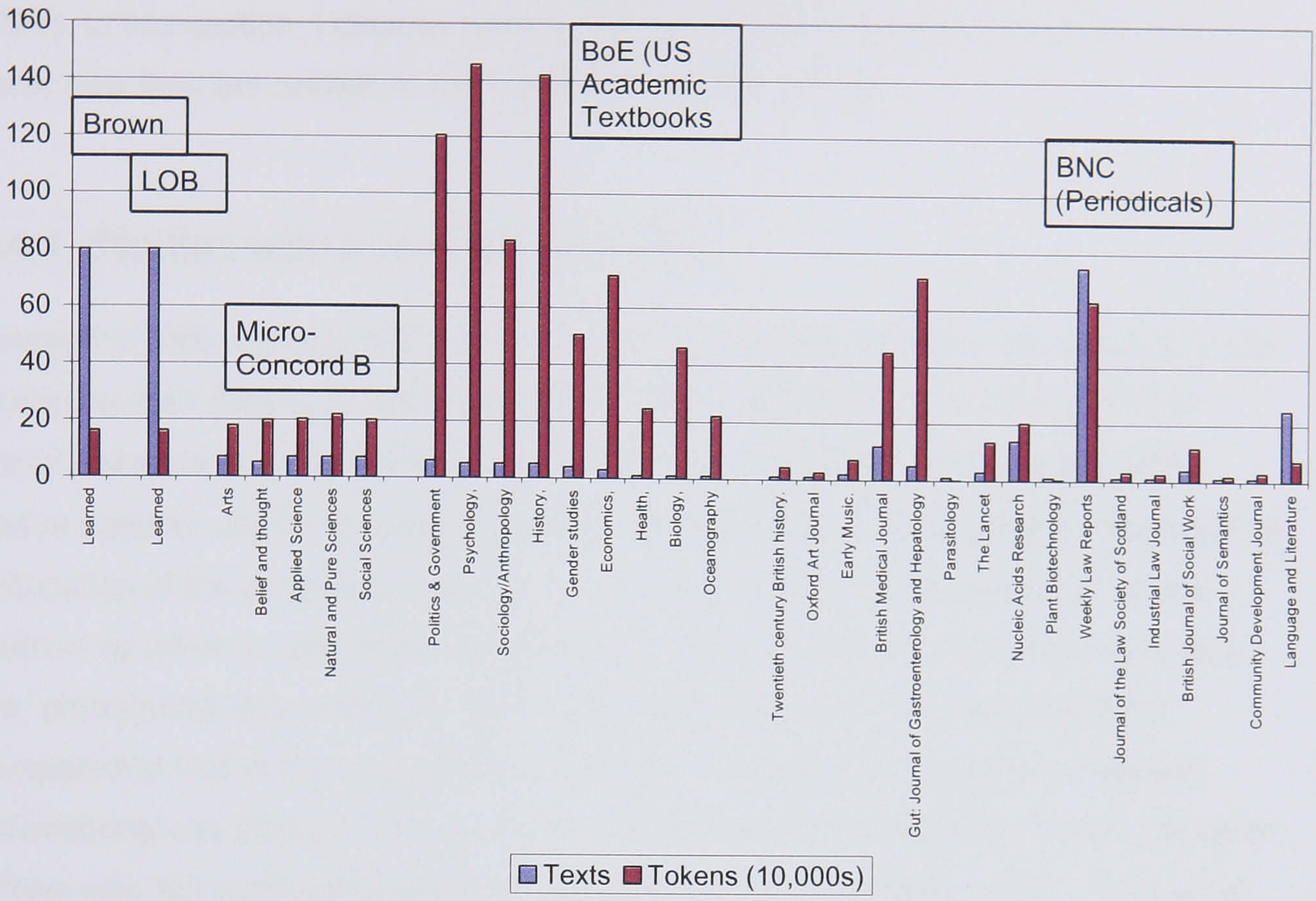


Figure 4.7: Comparison of academic sub-sections of Brown, LOB, MicroConcord B, Bank of English, and the British National Corpus

Having explained the decision to compile a corpus for the present study, I now draw together previous research on Phraseology, Corpus Linguistics, and Academic Discourse in a description of how the current study is similar and different to previous corpus-informed studies of word combinations in academic discourse.

4.4 Corpus studies of word combinations in academic discourse

As already discussed in section 2.3.2 above, the progression beyond the word as a useful unit of meaning (cf. Sinclair 1996, 1998; Danielsson 2007; Gardner 2007) has led to a great deal of research in what I have already collectively termed *extended empirical collocations*: multi-word phrases of varying lengths, degrees of fixedness and variability whose status as linguistic units is not yet determined. At the same time, more attention has been paid to the communicative conventions that characterise different contexts of situation. Also, as will be seen below, the linguistic features of registers and genres and the pragmatic features of communication within

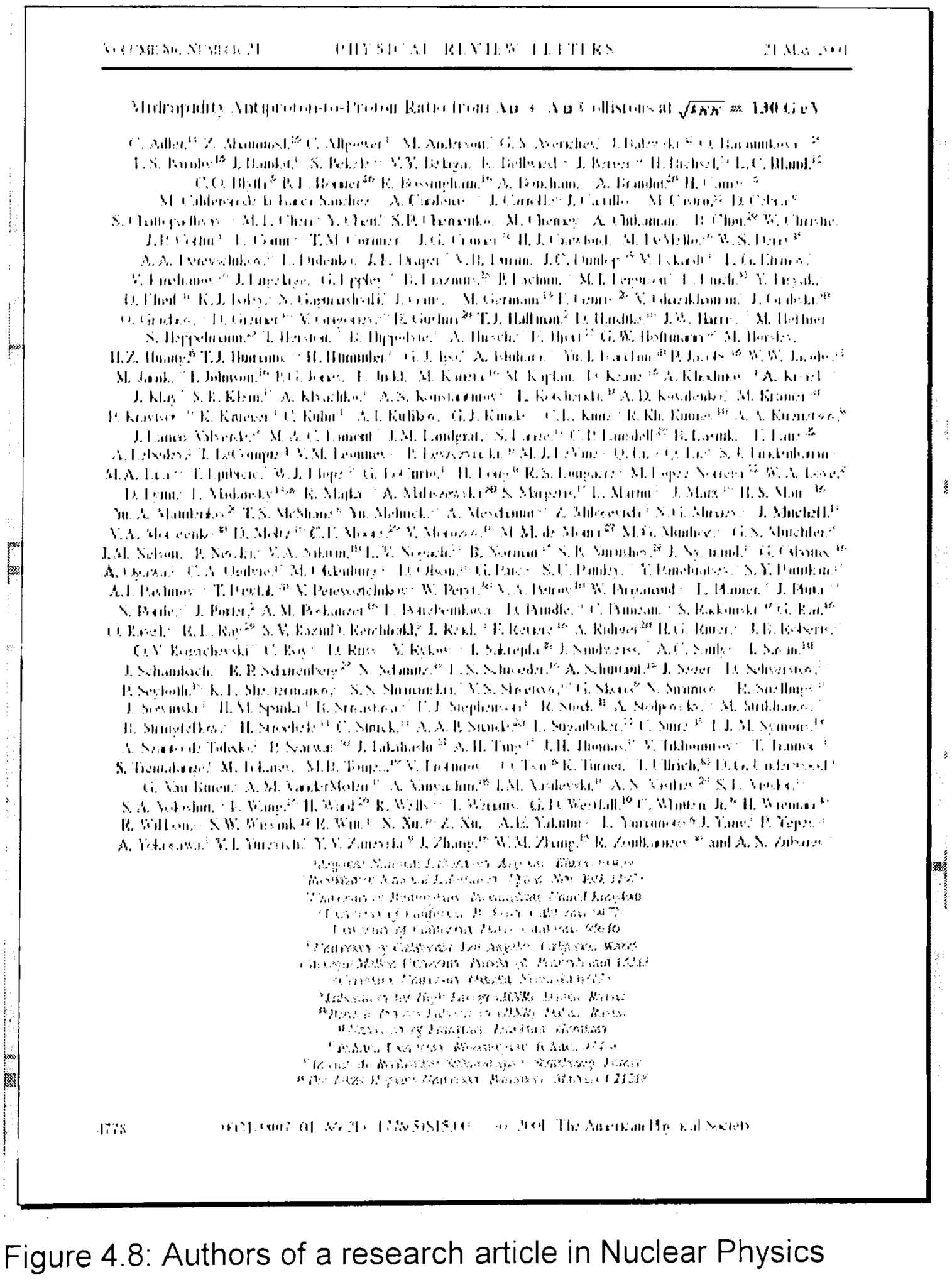
discourse communities have been described in more detail (Biber et al. 1998; Hyland 2000). In this section, I discuss work on some of these extended collocations, and show how they are related to the goals of the present study.

4.4.1 Studies with a pedagogical aim

Howarth (1996), working within the Eastern Phraseology tradition (see section 2.2.2), studied written data from native and non-native student writers of English in the social sciences to compare their use of restricted collocations (Aisenstadt 1981). Native speaker use is an important aspect of the Eastern Phraseological approach to collocation at the pragmatic level, since "acceptability to a native speaker" is one method by which a collocation is judged to be institutionalised. Howarth measured the "phraseologically deviant" (1996: 156) performance of non-native students compared to that of native students, but could find no direct connection between collocational use and proficiency. Some errors were in fact made by "more proficient writers who felt confident enough to experiment with collocability," (ibid.: 159) while writers felt to be less proficient seemed to avoid producing such features.

Howarth's work is an example of a corpus-based approach, which elucidates existing phraseological theory on restricted collocations from corpus evidence. As I have argued on the other hand, the present study does not seek to investigate the native/non-native dimension of fixed collocational patterns, since this is not a criterion for their identification. Occurrences in the data will all be attested, published instances of language use and will have already been accepted by the discourse community. As I have argued in section 2.4.3, whether a fixed collocational pattern is "right" or "wrong" is immaterial: this study investigates whether it occurs frequently in the data and how it is used. Secondly, identifying whether a writer of a particular text is a native or non-native speaker is very difficult. Contributors to international journals in English belong to a global academic community where the mother tongue is not a membership criterion. Peacock (2007) refers to a method of identifying native English speakers by Anglo-Saxon sounding names with an affiliation to an institution in a native-speaking country, but this risks miscategorising second or third generation migrants to native speaking countries.

The concept of "the author" itself in academic journal articles is also unclear. In some disciplines, articles are mostly co-authored and it is impossible to identify an individual author. In some disciplines journal articles are generally monographs, while in others, particularly when research from large-scale projects is reported, the number of listed authors is close to three figures (see Figure 4.8 for an example from Nuclear Physics). In recognition of this difficulty, in this thesis I more often refer to "how a pattern is used in a text" than "how the writer of a text uses a pattern."



VOLUME 86, NUMBER 21 PHYSICAL REVIEW LETTERS 21 MAY 2001

Midrapidity Antiproton-to-Proton Ratio from Au + Au Collisions at $\sqrt{s_{NN}} = 130$ GeV

C. Adler,[14] Z. Ahammed,[29] C. Allgower,[1] M. Anderson,[2] G. S. Averichev,[7] J. Balewski,[8] O. Barannikova,[7,29] L. S. Barnby,[16] J. Baudot,[9] S. Bekele,[17] V. V. Belaga,[7] R. Bellwied,[27] J. Berger,[11] H. Bichsel,[28] L. C. Bland,[2] C. O. Blyth,[3] B. E. Bonner,[30] R. Bossingham,[15] A. Boucham,[24] A. Brandin,[20] H. Caines,[4] M. Calderón de la Barca Sánchez,[2] A. Cardenas,[27] J. Carroll,[15] J. Castillo,[24] M. Castro,[27] D. Cebra,[5] S. Chattopadhyay,[29] M. L. Chen,[1] Y. Chen,[6] S. P. Chernenko,[7] M. Cherney,[10] A. Chikanian,[4] B. Choi,[25] W. Christie,[2] J. P. Coffin,[9] T. M. Cormier,[27] J. G. Cramer,[28] H. J. Crawford,[5] M. DeMello,[17] W. S. Deng,[15] A. A. Derevschikov,[18] L. Didenko,[2] J. E. Draper,[5] V. B. Dunin,[7] J. C. Dunlop,[2] V. Eckardt,[11] L. G. Efimov,[7] V. Emelianov,[20] J. Engelage,[5] G. Eppley,[23] B. Erazmus,[24] P. Fachini,[2] M. L. Fergason,[2] J. Fine,[2,29] Y. Fisyak,[2] D. Flierl,[11] K. J. Foley,[2] N. Gagunashvili,[7] J. Gans,[4] M. Germain,[9] F. Geurts,[23] V. Ghazikhanian,[6] J. Grabski,[26] O. Grachov,[27] D. Greiner,[13] V. Grigoriev,[20] E. Gushin,[20] T. J. Hallman,[2] D. Hardtke,[15] J. W. Harris,[4] M. Heffner,[5] S. Heppelmann,[22] T. Herston,[29] E. Hjort,[15] A. Hirsch,[18] E. Hjort,[15] G. W. Hoffmann,[25] M. Horsley,[4] H. Z. Huang,[6] T. J. Humanic,[17] H. Hümmler,[11] G. J. Igo,[6] A. Ishihara,[25] Yu. I. Ivanshin,[19] P. Jacobs,[15] W. W. Jacobs,[12] M. Janik,[26] I. Johnson,[15] P. G. Jones,[3] E. Judd,[2] M. Kaneta,[15] M. Kaplan,[16] D. Keane,[14] A. Khodinov,[20] A. Kisiel,[26] J. Klay,[15] S. R. Klein,[15] A. Klyachko,[12] A. S. Konstantinov,[18] L. Kotchenda,[20] A. D. Kovalenko,[7] M. Kramer,[21] P. Kravtsov,[20] K. Krueger,[1] C. Kuhn,[9] A. I. Kulikov,[7] G. J. Kunde,[4] C. L. Kunz,[12] R. Kh. Kutuev,[10] A. A. Kuznetsov,[7] J. Lamas Valverde,[17] M. A. C. Lamont,[2] J. M. Landgraf,[2] S. Lange,[11] C. P. Lansdell,[25] B. Lasiuk,[4] F. Laue,[2] A. Lebedev,[2] T. LeCompte,[1] V. M. Leontiev,[18] P. Liaw,[6] M. J. LeVine,[2] Q. Li,[27] Q. Li,[27] S. J. Lindenbaum,[21] M. A. Lisa,[17] T. Ljubicic,[2] W. J. Llope,[23] G. LoCurto,[11] H. Long,[6] R. S. Longacre,[2] M. Lopez-Noriega,[17] W. A. Love,[2] D. Lynn,[2] L. Madansky,[13,30] R. Majka,[4] A. Maksimchuk,[20] S. Margetis,[14] L. Martin,[9] J. Marx,[15] H. S. Matis,[15] Yu. A. Matulenko,[18] T. S. McShane,[10] Yu. Melnick,[18] A. Meschanin,[18] Z. Milosevich,[5] N. G. Minaev,[18] J. Mitchell,[23] V. A. Moiseenko,[19] D. Moltz,[15] C. F. Moore,[25] V. Morozov,[2] M. M. de Moura,[27] M. G. Munhoz,[17] G. S. Mutchler,[23] J. M. Nelson,[3] P. Nevski,[2] V. A. Nikitin,[11] L. V. Nogach,[18] B. Norman,[14] S. B. Nurushev,[18] J. Nystrand,[6] G. Odyniec,[15] A. Ogawa,[2] C. A. Ogilvie,[15] M. Oldenburg,[15] D. Olson,[15] G. Paic,[17] S. U. Pandey,[27] Y. Panebratsev,[7] S. Y. Panitkin,[2] A. I. Pavlinov,[27] T. Pawlak,[26] V. Perevoztchikov,[2] W. Peryt,[26] V. A. Petrov,[20] W. Pinganaud,[24] E. Platner,[2] J. Pluta,[26] N. Porile,[21] J. Porter,[2] A. M. Poskanzer,[15] E. Potrebenikova,[7] D. Prindle,[28] C. Pruneau,[27] S. Radomski,[23] G. Rai,[15] O. Ravel,[24] R. L. Ray,[25] S. V. Razin,[7,20] D. Reichhold,[21] J. Reid,[28] F. Retiere,[15] A. Ridiger,[20] H. G. Ritter,[15] J. B. Roberts,[23] O. V. Rogachevski,[7] C. Roy,[24] L. Ruan,[2] V. Rykov,[27] I. Sakrejda,[15] J. Sandweiss,[30] A. C. Saulys,[2] I. Savin,[8] J. Schambach,[25] R. P. Scharenberg,[21] N. Schmitz,[13] L. S. Schroeder,[15] A. Schüttauf,[13] J. Seger,[12] D. Seliverstov,[20] P. Seyboth,[13] K. E. Shestermanov,[18] S. S. Shimanskii,[7] V. S. Shirokov,[20] G. Skoro,[7] N. Smirnov,[30] R. Snellings,[15] J. Sowinski,[12] H. M. Spinka,[1] B. Srivastava,[21] T. J. Stephenson,[13] R. Stock,[11] A. Stolpovsky,[27] M. Strikhanov,[20] B. Stringfellow,[21] H. Stroebele,[11] C. Struck,[13] A. A. P. Suaide,[27] E. Sugarbaker,[17] C. Suire,[2] T. J. M. Symons,[15] A. Szanto de Toledo,[17] P. Szarwas,[26] J. Takahashi,[8] A. H. Tang,[14] J. H. Thomas,[15] V. Tikhomirov,[20] T. Trainor,[28] S. Trentalange,[6] M. Tokarev,[7] M. B. Tonjes,[22] V. Trofimov,[20] O. Tsai,[6] K. Turner,[2] T. Ullrich,[30] D. G. Underwood,[1] G. Van Buren,[2] A. M. VanderMolen,[12] A. Vanyashin,[20] I. M. Vasilevski,[8] A. N. Vasiliev,[18] S. E. Vigdor,[12] S. A. Voloshin,[27] F. Wang,[21] H. Ward,[25] R. Wells,[17] T. Wenaus,[2] G. D. Westfall,[10] C. Whitten Jr.,[6] H. Wieman,[15] R. Willson,[17] S. W. Wissink,[12] R. Witt,[1] N. Xu,[15] Z. Xu,[2] A. E. Yakutin,[18] I. Yamamoto,[15] J. Yang,[6] P. Yepes,[23] A. Yokosawa,[1] V. I. Yurevich,[7] Y. V. Zanevski,[7] J. Zhang,[15] W. M. Zhang,[12] R. Zoulkarneev,[8] and A. N. Zubarev[7]

[1]Argonne National Laboratory, Argonne, Illinois 60439
[2]Brookhaven National Laboratory, Upton, New York 11973
[3]University of Birmingham, Birmingham, United Kingdom
[4]University of California, Berkeley, California 94720
[5]University of California, Davis, California 95616
[6]University of California, Los Angeles, California 90095
[7]Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
[8]Creighton University, Omaha, Nebraska 68178
[9]Laboratory for High Energy (IReS), Strasbourg, France
[10]Michigan State University, East Lansing, Michigan 48824
[11]Max-Planck-Institut für Physik, Munich, Germany
[12]Nuclear Physics Laboratory, University of Washington, Seattle, Washington 98195
[13]Physics Department, University of Frankfurt, Frankfurt, Germany
[14]Purdue University, West Lafayette, Indiana 47907

 The American Physical Society

Figure 4.8: Authors of a research article in Nuclear Physics

A well-known pedagogical study which described extended collocations in academic writing was *Lexical Phrases and Language Teaching* (Nattinger & DeCarrico 1992). Nattinger and DeCarrico's construct, the *lexical phrase*, is a collocation, continuous or discontinuous, fixed or variable, which has a pragmatic discourse organising function. Nattinger and Decarrico presented a taxonomy of lexical phrases based on their form and function. For example, they describe a type of lexical phrase called a 'phrasal constraint' which could take a form like *the* ____*er the* ____*er*, which would structure a phrase like *the more the merrier.* These shorter lexical phrases, according to Nattinger and DeCarrico's model, can be built up into extended collocations. For example turning a phrasal constraint into a 'sentence builder' requires the addition of noun groups or verb groups as in *the* ____*er X, the* ____ *Y*, which occurs in phrases like *the harder they come, the harder they fall.* These sentence-building lexical phrases can be extended to produce quite complex frames offering paradigmatic choices of verbs and other optional clause elements, marked by brackets, as in Example 4.7:

> Example 4.7: *this paper will compare/contrast/describe/demonstrate that X (first) (by analysing/comparing/demonstrating (that Y) (then by _____ing Z, and finally by _____ing A*;

The lexical phrase was conceived as a unit with direct applicability to language teaching, and Nattinger and DeCarrico specified discourse models, such as those underlying informal and business letters or academic essays, together with the lexical phrases which they intended learners to use when writing these texts. This meant, however, that whether a collocation can be classed as a lexical phrase depends very much on the text in which it is used, since it must have the specified communicative function according to the discourse model for that text proposed by Nattinger and DeCarrico.

In the case of lexical phrases specified for use in written academic discourse, the model for academic discourse was the freshman composition. This is an essay written by undergraduate university students in the United States who are all expected to complete two composition courses in order to fulfil their English language and writing requirements for their degree programmes (Macksoud, personal communication). The "freshman comp." typically follows a five-paragraph

sequence of opening, body, and closing. The lexical phrases specified as typical of freshman comp. writing, therefore, were only suitable for teaching students who needed to produce such texts. It is not clear what the data was from which these lexical phrases were identified by Nattinger and DeCarrico, other than that it was "written discourse collected from a variety of textbooks for ESL, textbooks for academic courses, letters to the editor of various news publications, and personal correspondence" (Nattinger & DeCarrico 1992: xvi). Little more information was given, such as how many tokens or texts there were for each of these different contexts of situation.

In view of the unclear nature of the data on which the notion of lexical phrases was based, in an earlier corpus-based study I searched in off-the-shelf corpora of academic English, namely the MicroConcord B corpus and academic periodicals from the BNC (following Lee's genre re-classification), in order to discover whether the form of the lexical phrase *it has/is been* ____ *that X* could be seen in attested data, and also whether the form-function relationship they specified for this lexical phrase in student essays was replicated in the different genres of books and research articles (Oakey 1998, 2002a). In the latter study I found that extended collocations with the same form as this lexical phrase could be found in academic writing, although there was considerably more paradigmatic variation of the verb, and the form was observed to perform more than one observable function. There was also evidence of *cross-functioning* (Moon 1998: 21) where a lexical phrase performed more than one function at the same time.

It seemed that by proposing a form-function relationship for lexical phrases, Nattinger and DeCarrico had made their lexical phrases model too dependent on their own discourse model, and thus reduced the potential applicability of lexical phrases, certainly to other genres such as research articles, and possibly to the register of academic writing in general. Again, however, the functions I identified in Oakey (1998; 2002a) are themselves suspect, since the MicroConcord B and the BNC are normative corpora containing text fragments which were not designed for contrastive studies such as these. The comparison of the present study results with those of Biber's and Hyland's lexical bundles, therefore, will hopefully reveal more about the functions of fixed collocational patterns in research articles.

4.4.2 Studies of disciplinary differences

Earlier work by Hyland (1996b; 1998b; 1998a) on pragmatics is concerned primarily with explaining the conventions of research articles (Hyland 2000: 90). He takes a corpus-based approach, in which corpora "assist not drive" (ibid.: 92) the research, in order to explore the use of *metadiscourse*. This is a theoretical concept from the field of pragmatics which is aligned with the Hallidayan interpersonal metafunction. It involves the evaluation of statements within a text, the creation of stance and the construction of discourse, and which functions rather like the "participant-oriented" lexical bundles discussed above. Hyland's corpus contained research articles from eight different disciplines and he searched for target functions (ibid.: 91) by using 180 items compiled from "grammars, dictionaries and earlier studies," (Hyland 1998a: 355).

Some of Hyland's metadiscourse markers are extended collocations, which he calls *boosters and hedges,* which construct academic discourse and show the writer's stance towards the knowledge they are creating, allowing

> writers to express conviction and assert a proposition with confidence, representing a strong claim about a state of affairs. Affectively they also mark involvement and solidarity with an audience, stressing shared information, group membership, and direct engagement, with readers
>
> (Hyland 1998a: 350)

For example, from the discussion section of a biology paper (ibid.: 377):

> Example 4.8: Recently it has been shown that over expression of active GAD causes severe abnormalities in growth and development of tobacco plants...
>
> Example 4.9: It is generally believed that the decline in cytoisolic pH during hypoxia might trigger the GAD activity that leads to GABA accumulation...

These two examples, according to Hyland, show that the writer uses these boosters to indicate that he accepts as facts these findings from other work. In his work Hyland continually refers to the "writers" of the texts he analyses. As I argued in section 4.4.1 above, however, it is difficult to confer a distinct identity on the "writer" of a research article submitted with more than 70 authors' names. This study will instead, as already stated, refer to *how language is used in a text* rather than to *how the writer uses language.*

Hyland's work on boosters and hedges is corpus-based: it reinforces existing notions on the pragmatics in journal articles but does specify the frequency of these forms in relation to each other. While the present study is more corpus-driven, as the fixed collocational patterns to be investigated will emerge from the data, it will employ Hyland's approach to corpus building, as it relied on non-linguists' opinions to determine which texts should be included. He interviewed practitioners in each academic discipline to find the journals which were most prestigious in each discipline, and which were most widely read. As far as possible research articles from the same journals as Hyland will be used.

Two corpus-driven studies of phraseology in academic discourse have shared a similar methodology, identifying statistically significant grammatical words and then examining their frequent collocational environments. Gledhill (Gledhill 2000a) looks at the collocation patterns of grammatical words in approximately 500,000 words from cancer research articles. Gledhill states that "analysis of grammatical words is an efficient way of arriving at a description of the most typical expressions in the corpus" (Gledhill 2000b: 117). He searches his corpus using words such as *of, for, has, is,* and *been,* before trying to identify the patterns in which they most frequently occur. He finds that *has been* is used with "extraposed that-clauses" (i.e. strings such as *it has been thought that*) which "project a research idea or fact through an empty subject" (ibid.: 121). He also finds that for signalling the general research aims of the article only *shown, found* and *demonstrated* can be used in both *it is/has been* ___ *that* forms, so a relationship between form, function, and genre can be seen in his work.

Groom (2007), in a contrastive study, takes a similar approach and compares two genres, those of the research article and book review, in Historical Studies and Art History. He is more concerned with the relationship between phraseology of grammatically salient words and the epistemology of different disciplines. Groom abstracts the frequent patterns into "semantic sequences" such as PROPERTY + of + ENTITY, realised, for example, in the noun group *major patterns <u>of</u> societal development*.

4.4.3 Corpus background to lexical bundles

The word combinations most similar to the fixed collocational patterns in this study are *lexical bundles* (Biber et al. 1999). Having discussed the functional frameworks above, I focus here more on the corpora used for their identification. As previously mentioned, lexical bundles are frequently recurring fixed strings of three or more words which "commonly go together in natural discourse" (ibid.: 990) and are similar in form to *n-grams* in section 2.3.2. Because computers process text as a progression of strings of characters and spaces, n-grams can include words from different sides of grammatical boundaries, or across turns. Lexical bundles, however, exclude strings which cross turns or contain punctuation. In order for a string to be classed as a lexical bundle, it must occur a certain number of times per million words in five or more texts in the corpus. Lexical bundles are therefore fixed collocational patterns, identified irrespective of their meaning, use, or the word classes their constituent words might belong to. Examples of lexical bundles are *one of the*, *part of the*, *the use of*, *in contrast to the* and *on the other hand*.

The aim of Biber et al. was to use lexical bundles to explore differences between spoken and written registers. The corpora representing the registers of academic prose and conversation for the first discussion of lexical bundles, part of the Longman Spoken and Written English Corpus (LSWE) (ibid.: 24), are shown in Figure 4.9 below. The corpus of academic prose contains 75 book extracts and 333 research articles, "nearly all" (ibid.: 33) of which are complete. The focus in that study was on comparing the forms and structures of lexical bundles found in the register of academic prose with those found in other registers such as conversation. Lexical bundles in conversation were found predominantly to begin with a pronominal subject followed by a verb phrase plus the start of a complement clause, such as *I don't know why* and *I thought that was,* while in academic prose the most common form was part of a noun phrase, as in *the nature of the,* or prepositional phrases such as *as a result of* (ibid.: 991).

It can be seen that Biber's two register subcorpora in which lexical bundles were first identified are not equal in size. The academic prose subcorpus contains around 1.4 million more tokens than the conversation subcorpus, but contains around 3,000 fewer texts. According to Biber et al. there was "no need to use subcorpora with

exactly the same number of words, because all frequency counts are normed to a common basis (occurrences per million words of text)" (ibid.: 25-26). A lexical bundle which is found to occur 50 times in 5 million tokens of academic discourse, for example, can be normed to a frequency of 10 times per million words, which would allow it to be compared with a lexical bundle which occurs 40 times in 4 million tokens of conversation. Once their frequency counts have been thus normalised, lexical bundles can therefore be compared across registers. Biber et al. say little, however, about how this disparity in text lengths, and the fragmented nature of some of these texts, might influence the lexical bundles they identify in their study.

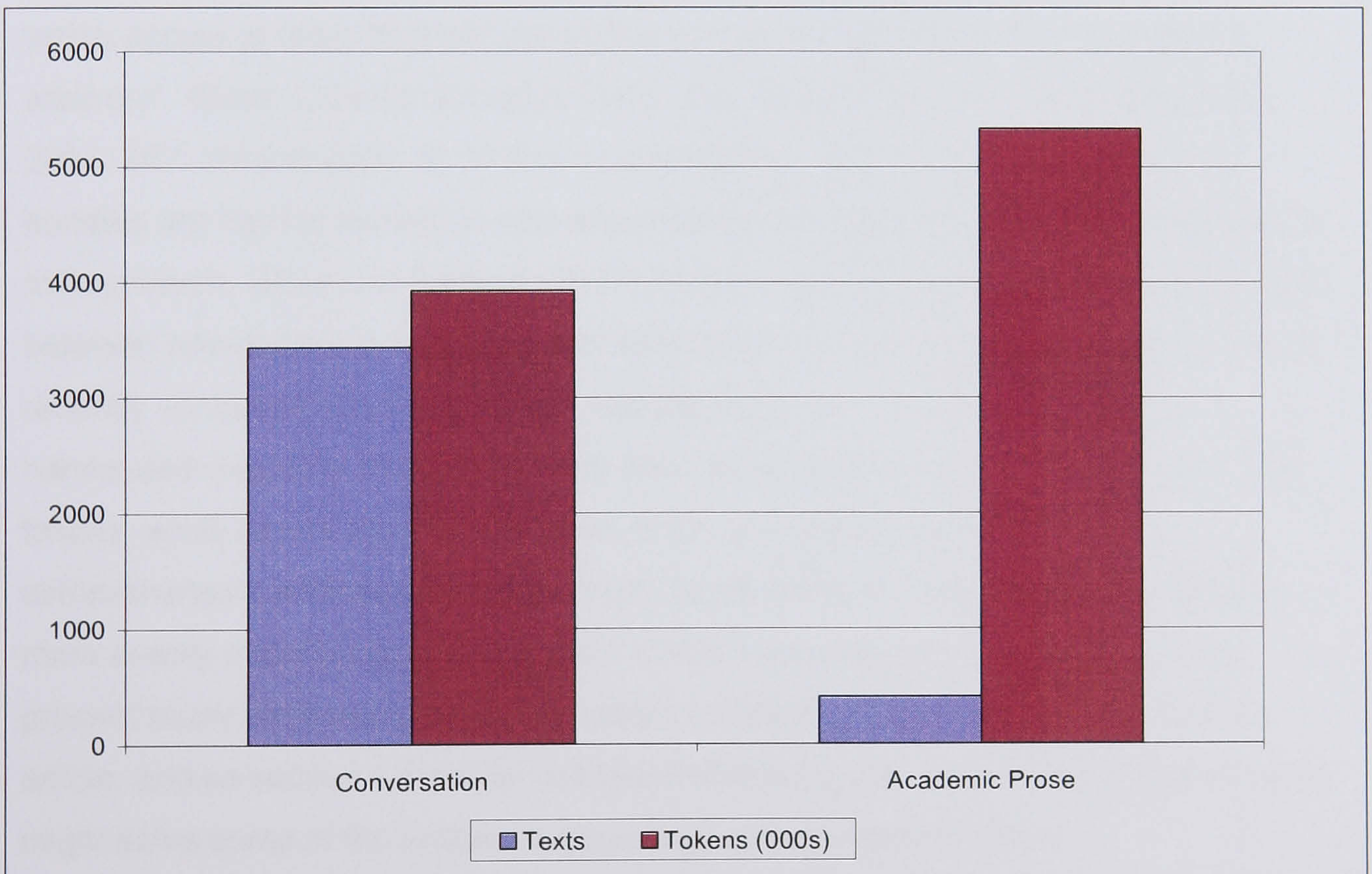


Figure 4.9: Texts and tokens in two of the subcorpora in Biber et al. (1999: 25)

We now come to the areas of uncertainty which the main research questions of this thesis attempt to clarify. The first problem stems from attempting to contrast discourse functions of lexical bundles in different registers. Grouping texts within a sample of register means that the sample often contains different numbers of texts from different genres. This in turn means that the register samples may be of different sizes, either in terms of the number of tokens or in terms of the numbers of texts, or in terms of both. The difficulty presented by unequal numbers of words is

traditionally overcome by normalising the results so that the frequency of occurrence is expressed in the same number of words, such as occurrences per million. This might be called an isolexical (from 'same number of words') comparison. Normalisation is possible when abstract linguistic categories such as passive verbs are being counted, for example, since these features are linearly distributed (Biber 2006: 253). Words, on the other hand follow a non-linear distribution, the 'long tail' already discussed. As argued above, this non-linear distribution means that normalising up or down to achieve an isolexical comparison risks skewing the results.

The lexical bundle is an isolexical construct, defined as an extended collocation which occurs at least 40 times per million words; the cut-off point is "somewhat arbitrary" (Biber, Conrad, & Cortes 2004: 376; Biber 2006: 134; Biber & Barbieri 2007: 267; Hyland 2008: 8). Without normalisation of results, therefore, lexical bundles are hard to identify in non-isolexical subcorpora which contain fewer than a million words. Biber and Barbieri (2007) in their comparative study of lexical bundles between non-isolexical subcorpora of spoken and written academic registers, have recently conceded that this problem has the potential to affect the results of normalised isolexical studies: some of their subcorpora contain more than a million tokens, while others are much smaller. Biber and Barbieri admit that "more comprehensive analyses based on much larger samples, with the sample design more evenly matched across registers" (268ff.) are required. The corpus in the present study attempts to meet this need in relation to the genre of the research article, and so section 4.5 below outlines how having two versions of a larger corpus might solve some of the problems caused by uneven sample sizes.

In addition to his contrastive work on register, Biber has also compared lexical bundles within a single genre, that of the academic textbook (2006: 164), from the textbook section of the T2K-SWAL corpus, comprising the disciplines of Business, Engineering, Natural Science, Social Science, and Humanities. The corpus is neither isolexical nor isotextual, and each subcorpus (see Figure 4.10 below) contains fewer than a million words, and so Biber adjusts the frequency cut-off point (to 20 occurrences per million words) and the minimum number of texts in which a lexical bundle occurs (either 3 or 2) in order to obtain sufficient results for a comparison.

Biber (2006: 166-168) then counted how many lexical bundle types there were with particular discourse functions each subcorpus, gaining the results shown in Figure 4.11 below. It can be seen that the number of bundles does not match the size of the subcorpus. While Engineering has the fewest tokens, Humanities has the fewest bundles. While Social Science has the most tokens, Natural Science has the most bundles:

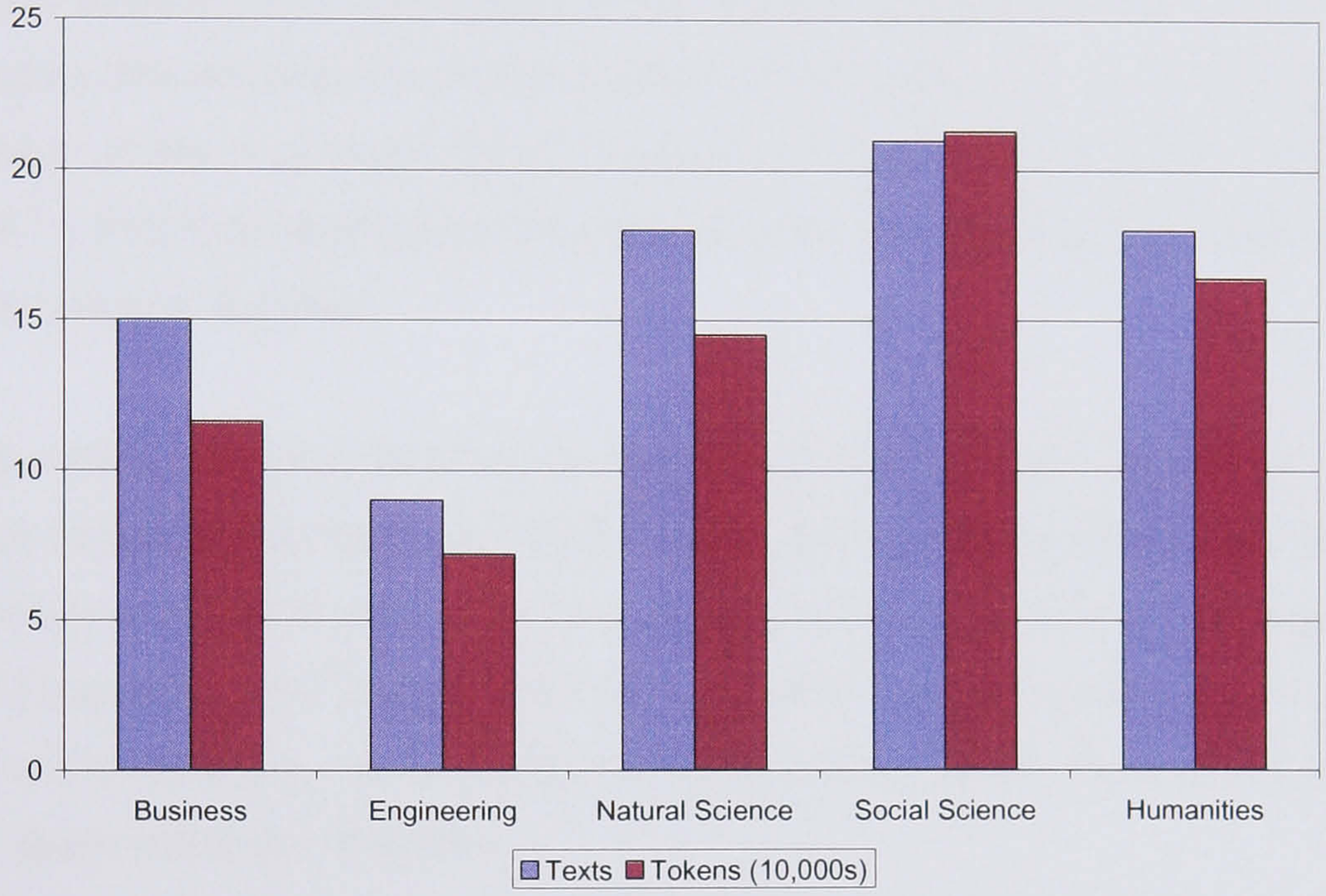


Figure 4.10: Texts and tokens in the textbook genre in Biber (2006: 25)

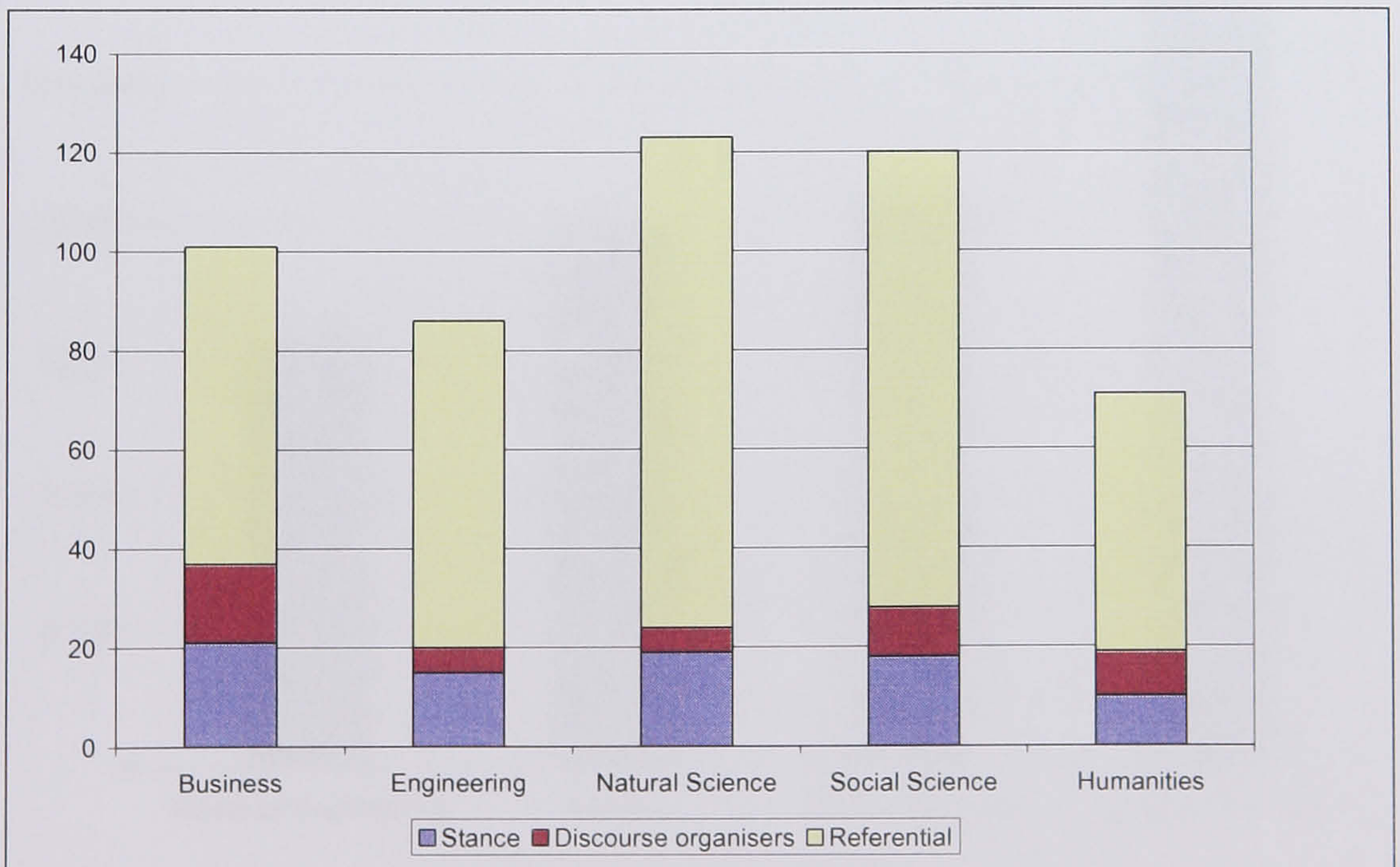


Figure 4.11: Distribution of lexical bundles across functional categories in academic textbooks (Biber 2006: 164)

While Biber suggests that the above results "need to be interpreted with caution, since the distributional patterns seen in the above figure do not in anyway correlate with sample size", he also adds that "we can be fairly confident that these are genuine differences rather than a mere artefact of sampling differences" (ibid.). I would suggest that this statement needs further investigation, which this study will be able to provide. Biber can only say that the number of bundle types in each subcorpus appears not to be related to the number of tokens in each subcorpus. If the samples are compared in terms of the number of texts i.e. in an isotextual comparison, there may indeed be a correlation. For example, it can be seen from Figure 4.11 that Engineering has the fewest texts, and also has the fewest discourse organising lexical bundles.

The other study of lexical bundles by Hyland (2008) is a register-based study of the functions of lexical bundles in academic prose, from a corpus of research articles, dissertations and MA theses in the disciplines of Biology, Electrical Engineering, Applied Linguistics, and Business Studies. Hyland does not specify the numbers of texts in each discipline, nor how many texts from each genre there are in total, but instead gives totals per discipline:

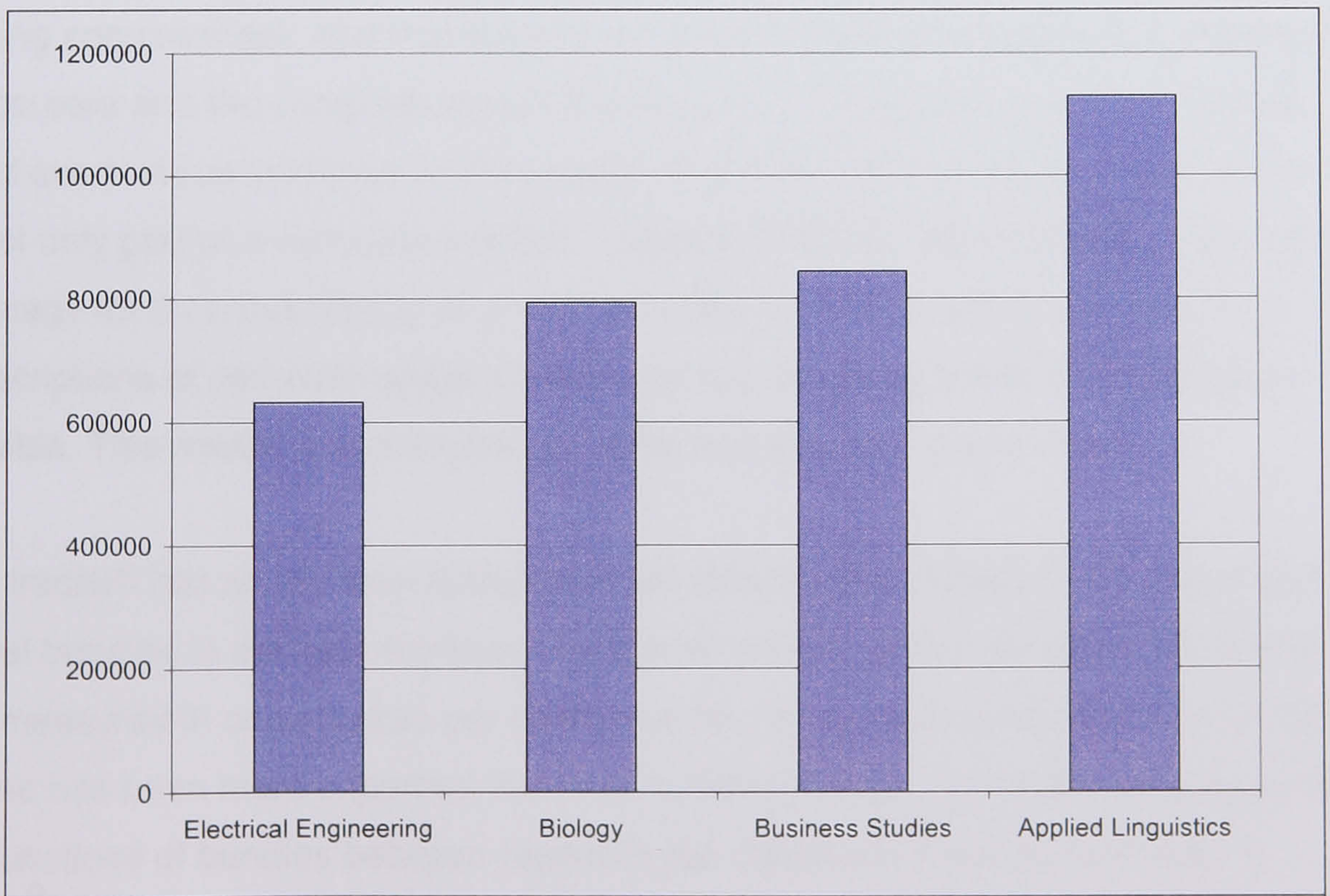


Figure 4.12: Texts and tokens in academic discourse (Hyland 2008)

Only one discipline subcorpus, that of Applied Linguistics, contains more than a million tokens, and so results for lexical bundles in the other subcorpora will require norming upwards in order to be quoted in terms of occurrences per million words, and so Hyland's results might also be skewed.

Problems seem to occur, therefore, when comparing lexical bundles between disciplines both in a single genre or within a single register. These are due to the inclusion of incomplete samples or the result of normalising upwards between subcorpora of different sizes. The means of addressing these problems in the present study is to propose a distinction between isolexical and isotextual corpora. This will be discussed further in the next section.

4.5 The advantages of a distinction between isolexical and isotextual versions of a comparative corpus

The difficulty in collecting a comparable amount of data needs to be addressed. When comparative subcorpora are compared, there is usually an imbalance between either the number of tokens in each subcorpus or the number of texts. Sinclair points out that balancing for an equal number of tokens is difficult to achieve without making compromises, and that there is a trade-off between the size of component subcorpora and the completeness of the texts in it. "Long texts in a small corpus could exert undue influence on the results of queries yet it is not good practice to select only part of a complete artefact," (Sinclair 2005: 6), and therefore, "samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size" (ibid.: 7).

This tradeoff has so far been acceptable for comparative studies of the *structures* of lexical bundles in different registers mentioned above. Lexical bundles are defined and measured in occurrences per million words, so in previous studies the number of tokens has been more important than the number of texts. However, comparisons of the *functions* of bundles between registers are potentially more problematic. A corpus of research article data with equal numbers of tokens in different disciplines but with different numbers of texts is only balanced for the amount of language that it

contains; it is not evenly balanced for the amount of communication that it contains. If each discipline contains a million tokens each, it will address the norming problem raised by Biber and Barbieri, but if there are different numbers of texts in each discipline then there are different numbers of communicative acts or situations in the corpus. In the genre of the research article there would be an imbalance in the number of genre stages: despite the fact that the subcorpora are evenly balanced in terms of the number of tokens, they contain different numbers of introductions, conclusions, and so on which might bias any results from the corpus in favour of the discipline with the largest number of texts.

For studies of the function of fixed collocational patterns, i.e. their possible roles in the act of communication, this possible bias can be investigated by examining two versions of the corpus. The isolexical corpus of research articles can be collected containing an equal number of tokens from each discipline to be compared. In this case it is possible to collect around a million words from each journal so that no norming will be required. The results from this corpus can be compared with the results from previous isolexical studies of lexical bundles between registers, as all results can be normalised to the same "occurrences per million words" measure which is the criterion for the identification of a lexical bundle. A better suggestion of the frequencies of the functions of fixed collocational patterns in texts, however, can be gained by creating another, isotextual, version of the corpus containing an equal number of texts. In this case a figure such as 100 texts per journal could be compared. The two versions would overlap: a certain number of texts would be included in order for a million tokens to be collected, and a certain number of tokens would be included in order for 100 texts to be collected, but there would be some texts in each version which were not included in the other.

What is not known is whether the fixed collocational patterns which are frequent in the isolexical version of a corpus would be the same fixed collocational patterns which are frequent in the isotextual version of the corpus. The re-balancing from equal numbers of tokens to equal numbers of texts will reduce the number of texts in some subcorpora but increase the number of texts in others. This changes the number of communicative acts in each subcorpus, and thus creates the potential for other fixed collocational patterns to become more frequent isotextually than they are

isolexically. To my knowledge this type of distinction has not been made before. Previous corpora have either been isolexical or isotextual, but not both.

Having investigated the two versions of the corpus, this study will then be able to clarify Biber's and Hyland's findings by introducing a distinction between constructs which are isolexical and constructs which are isotextual. The lexical bundle is an isolexical construct because its frequency in different registers has been measured in terms of the same number of tokens (per million words) so that its frequency can be shown in terms of the same amount of language in each register. The "fixed collocational pattern" in this study will be an isotextual construct with a number of occurrences per communicative event, and thus should be a better indicator of the relative amounts of functions in the same genre.

The decision of which disciplines to include is based on Hyland's Journal Corpus (1998a) which he used to investigate the boosters and hedges discussed above. This corpus consists of eight independent subcorpora, four from Science disciplines and four from non-Science disciplines (see Table 4.1 below). Since Hyland was looking at discourse functions of these items, his corpus was isotextual. He included one article each from seven journals in eight disciplines, totalling 330,000 tokens. He does not specify the number of tokens in each discipline, as his study did not seek to contrast the frequency of occurrence of boosters and hedges in different disciplines.

Discipline	**Texts**	**Tokens**
Microbiology	7	not specified
Electrical Engineering	7	not specified
Mechanical Engineering	7	not specified
Physics	7	not specified
Applied Linguistics	7	not specified
Marketing	7	not specified
Philosophy	7	not specified
Sociology	7	not specified
Total	**56**	**330k**

Table 4.1: Distribution of texts in the Journal Corpus (Hyland 1998a)

This section has established the main focus of the study, namely that fixed collocations can be viewed as isolexical constructs such as lexical bundles, and also can be viewed as isotextual constructs. The isolexical perspective highlights occurrences in language, whereas the isotextual perspective highlights occurrences in communication. The first aim of the study therefore is to compare isolexical occurrences of fixed collocational patterns across disciplines, and compare isotextual occurrences of fixed collocational patterns across disciplines, and then compare the isolexical occurrences with isotextual occurrences. These results can then be compared with the results of the non-isotextual studies in Biber (2006) and Hyland (2008) in order to reveal more about the register and genre-based behaviour of lexical bundles. We have now therefore come to the stage in this thesis where its research questions can be re-stated in more detail.

4.6 Detailed research questions

The distinction between the lexical bundle as an isolexical construct and the fixed collocational pattern as an isotextual construct requires a refinement of the initial research questions. The first question addressed in this study is to find the most frequent fixed collocational patterns in a sample of research papers in different academic disciplines, but the samples in each discipline must be comparable. This study therefore should also compare occurrences of fixed collocational patterns between the isolexical and isotextual versions of the corpus. It is hoped that this will reveal whether the imbalance between the number of texts and tokens in each subcorpus influences which patterns are frequent. The first research question can accordingly be re-expressed as:

1. What are the most frequent fixed collocational patterns in

a) an isolexical sample of an equal number of tokens from journal articles from eight academic disciplines?

b) an isotextual sample of an equal number of texts from journal articles from eight academic disciplines?

The second question, which attempts to identify fixed collocations across disciplines, now also needs to compare across disciplines isolexically and also isotextually. Again, this will hopefully reveal whether the imbalance between the number of texts and tokens in each subcorpus influences which patterns are frequent. The answer to this question will require comparisons not only between Science and non-Science disciplines, but also between isolexical disciplines and isotextual disciplines. The second research question can accordingly be re-expressed as:

2. Do any fixed collocational patterns occur frequently in all eight academic disciplines in
a) an isolexical sample of an equal number of tokens from journal articles from eight academic disciplines?
b) an isotextual sample of an equal number of texts from journal articles from eight academic disciplines?

The third question will need to be focused on investigating whether the isotextual construct identified in the first two questions appears in the isolexical results of Biber (2006) and Hyland (2008). If any fixed collocational patterns are identified by the first two questions, then the isotextual corpus can be investigated to see whether these patterns perform the discourse functions proposed for them by Biber and by Hyland. These functions will be genre-based, rather than register-based, and so the location of each pattern in the text in which it is used will be investigated alongside its discourse function. The third research question can accordingly be re-expressed as:

3. Are any of the fixed collocational patterns identified in 1 and 2 found in Biber (2006) and Hyland (2008)? If so, is there a relationship between textual environment, namely at the syntactic, semantic, and pragmatic levels, of these fixed collocational patterns
a) within disciplines?
b) across disciplines?

4.7 Summary

Having looked at previous work on aspects of Phraseology, Academic Discourse analysis, and Corpus Linguistics, and having refined the research questions, we now turn to methodological considerations. It is now possible for the contrastive corpus of journal articles to be used in this study to be larger than those previously constructed. In the same way that the boom in language data in the 1980s permitted the construction of academic-industrial corpora, the development of the internet in the late 1990s has given individual researchers access to large amounts of data from a wide variety of language contexts. This easier access, coupled with the powerful computing power available in most desktops and university computer clusters, has allowed larger corpora to be collected and studied by individual scholars working autonomously. The corpus in the present study takes advantage of these developments, and the next chapter will describe its overall design and construction, how comparable subcorpora were created, and the methods used to tackle the research questions.

Chapter 5 Methodology: working with corpora

5.1 Introduction

This chapter details the design and construction of the research article corpus used in this study. It presents quantitative information about the size and composition of its isolexical and isotextual versions. The second section describes the corpus-driven approach to be used to identify fixed collocational patterns in both versions in order to answer the first two research questions. The final section describes the methodology to be used in analysing any fixed collocational patterns which are identified in relation to the third research question.

5.2 The design and construction of the corpus in this study

5.2.1 The design of the corpus

The corpus in this study is contrastive, and comprises data from research articles in different academic disciplines. The subcorpus for each discipline is intended to be similar enough for it to be compared with the subcorpora for other disciplines without any normalisation. The corpus is designed to follow the composition of Hyland's 1998 Journal Corpus. Because the collocational patterns in this corpus-driven study emerge from the data, the corpus has to be sufficiently large for any frequently occurring patterns to be noticeably more frequent than other patterns. Studies of smaller corpora can identify differences between disciplines, but if the differences in frequency are only one or two, then this is not enough to illustrate clear differences. Normalised occurrences per million words of different permutations of *it is/has been* _____ *that* barely pass single figures in the MicroConcord B corpus and the periodicals section of the BNC (Oakey 1998, 2002a).

I therefore decided to collect a corpus similar in design to Hyland's corpus, but on a larger scale. For the size of the isolexical version of the corpus, a figure of a million words per journal was chosen. This meant that isolexical frequencies would be per five million words, and comparisons of fixed collocational patterns with lexical bundles could be made without the need for any normalisation.

The disciplines represented in the corpus were mainly the same as in Hyland's corpus, but two of the disciplines chosen were different. Physics at my institution largely consists of Nuclear Physics, as the department has been active in teaching and research in the subject since the 1930s, and so the more specialised discipline was substituted. The discipline of Applied Linguistics was also dropped from the corpus, since it was felt that it was too close to the subject of this study. The potential existed for me to say something about the language of Applied Linguistics based on my insider's knowledge of the subject which I could not say for the language in any of the other disciplines. The discipline of Economics was substituted in its place in order to maintain the basic Science/Non-Science balance of the subcorpora.

Following Hyland's methodology, five journals per discipline were used: the journals chosen in each discipline were the same as Hyland's study or identified by the relevant head of department at the University of Birmingham as being the most prestigious in each discipline. My request for guidance and the responses it elicited from the various heads of department can be found in Appendix 1.

5.2.2 The corpus construction procedure

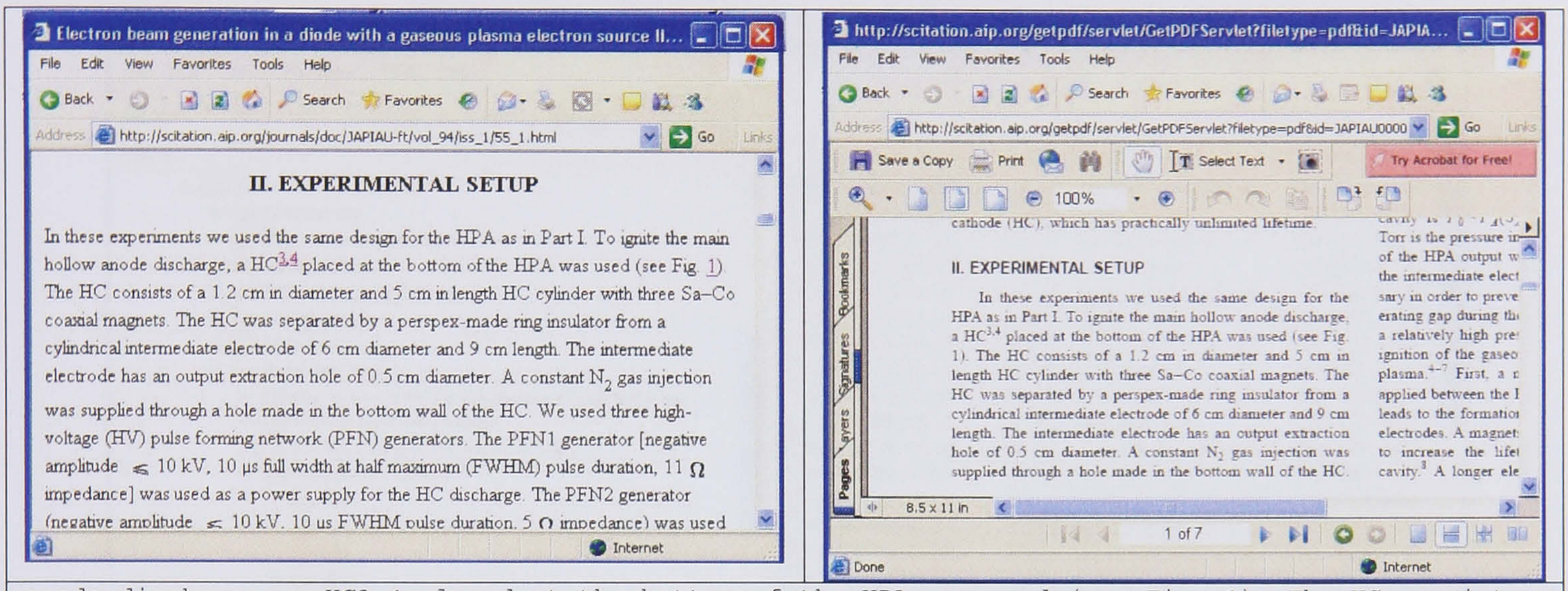
The creation by one researcher of a corpus this size - over twice as large as the original Bank of English - was only possible due to the many technological innovations which have been applied to the production, dissemination, and preservation of academic texts since the 1970s. Each of these technological innovations has quickly been superseded by newer technology; the microfiche of the 1960s and 1970s, in which very small graphical reproductions of texts were stored on plastic sheets, were in the 1980s replaced by digital reproductions of texts which were stored on compact discs and read using computer software. These CDROMs became the storage medium of choice for many publishers of academic journals, and university libraries introduced networked computers on which library users could browse the back-issues held centrally on CDROM. This model of information storage and retrieval was in its turn quickly outmoded, however, by the introduction of networked servers which held as much information as thousands of CDROMs, and in the late 1990s these CDROM journals began to be replaced by subscription services to the internet. University libraries in the UK have continued to encourage

library users to make use of online electronic journals in order to free up shelf space for clusters of PCs on which these electronic journals can be read.

The terms of use of these subscription services permit downloading and storing of articles for individual scholarly study, and - provided a user does not download the whole, i.e. every single part, of a particular issue - constructing an electronic collection of journal articles in the form of a corpus is permitted. The research articles in the corpus can only be used for my personal study and will not be available to other researchers. This is unfortunate, but is a condition of being allowed to collect the data in the first place. The University of Birmingham's online journals service is based on software called Metalib, supplied by Ex Libris Ltd, and known as the eLibrary. At the time when the corpus for this study was collected, the University of Birmingham was using the TDNet electronic journals management system produced by TDNet Ltd, a subsidiary of Teldan Information Systems.

The corpus was created by downloading journal articles from the eLibrary and preparing them for use by *Wordsmith Tools 4* lexical analysis software (details of which are in section 5.3 below). This software processes text in plain text files, i.e. the simplest form of electronic text storage which contain characters stored as a series of 1s and 0s, and which has a .txt file extension. However, most online journal articles are only available in hypertext markup language (html) or portable document format (pdf), which allow extra information about the text to be displayed, such as fonts, graphics, and formulae. It is therefore necessary to convert these into plain text. Conversion is usually a straightforward process, since files in both pdf and html contain plain text stored alongside the other information. This plain text can in most cases be electronically selected, copied, and pasted into plain text files.

Figure 5.1 below shows the same portion of a research article displayed in the three different formats:




```
anode discharge, a HC3,4 placed at the bottom of the HPA was used (see Fig. 1). The HC consists
of a 1.2 cm in diameter and 5 cm in length HC cylinder with three Sa-Co coaxial magnets. The HC
was separated by a perspex-made ring insulator from a cylindrical intermediate electrode of 6 cm
diameter and 9 cm length. The intermediate electrode has an output extraction hole of 0.5 cm
diameter. A constant N2 gas injection was supplied through a hole made in the bottom wall of the
HC. We used three high-voltage (HV) pulse forming network (PFN) generators. The PFN1 generator
[negative amplitude 10 kV, 10 μs full width at half maximum (FWHM) pulse duration, 11
impedance] was used as a power supply for the HC discharge. The PFN2 generator (negative
amplitude 10 kV, 10 μs FWHM pulse duration, 5  impedance) was
```

Figure 5.1: Part of a journal article in pdf format (above right) html format (above left), and plain text format (above)

Some decisions on which journals to include in the corpus were subject to technological restraints. Journals which publish articles in protected pdf format could not be included in the corpus, as text cannot be extracted directly from such files. A more time-consuming procedure is required, in which the articles are either printed onto paper and then scanned into text files, or saved in a document image file (with a .tiff file extension) from which the text can be extracted using optical character recognition software. The one discipline affected by this type of problem was Mechanical Engineering, for which only one of the most prestigious journals was available electronically. Other international journals cited by members of the department were substituted.

Once downloaded, two copies of each journal article file were stored on the hard disk drive of a PC in a hierarchical folder structure, as on the left in Figure 5.2 below. The version of the article in plain text format was the one processed by *Wordsmith Tools*, while the version in pdf format, where available, allowed the pattern to be examined in its original typographical context in the qualitative stage of the analysis (see 5.5.4 below).

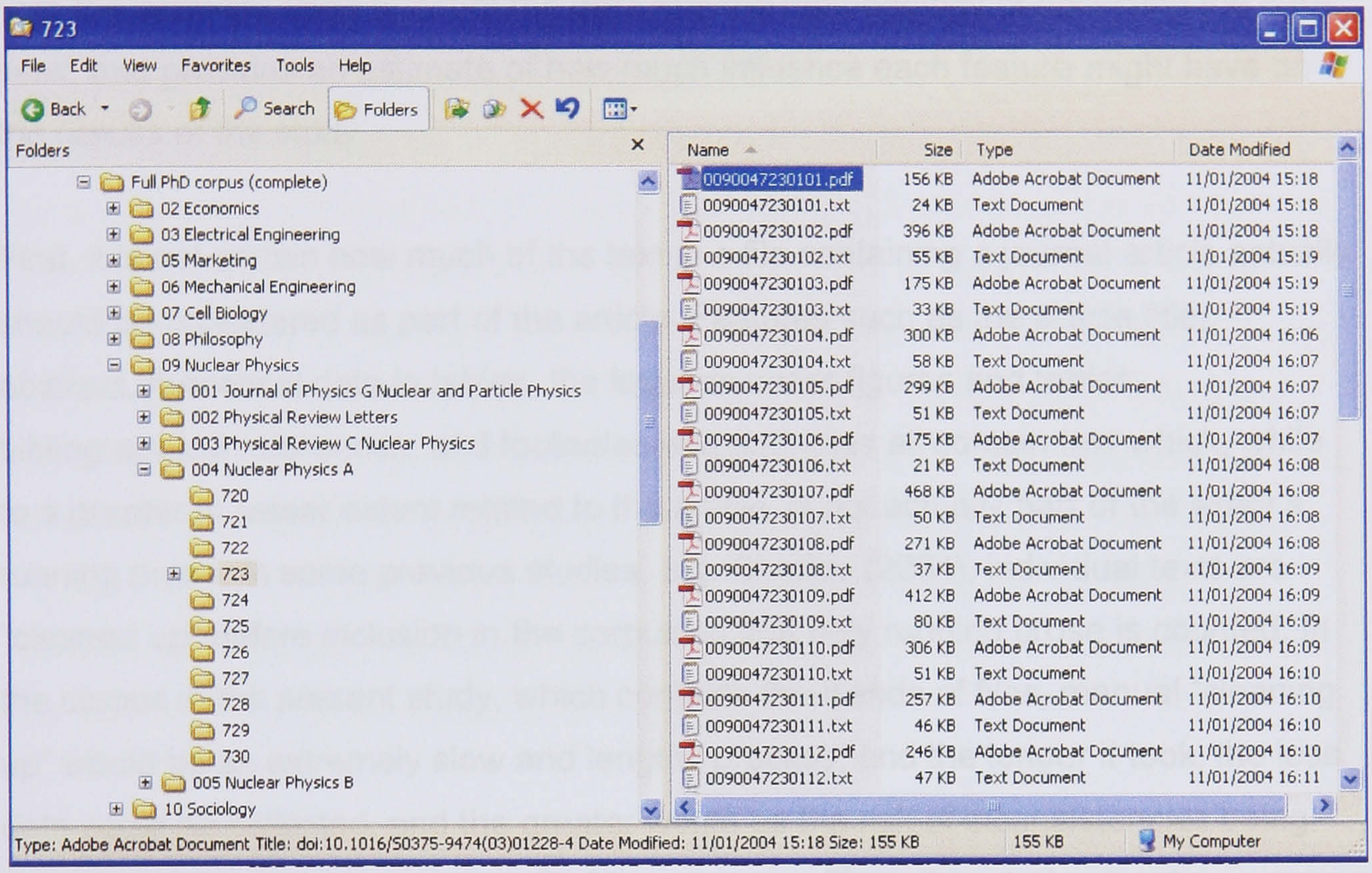


Figure 5.2: Folder structure for storage of the research article corpus

A simple 13-digit file-naming system was devised for easy identification of the files in the corpus, as on the right in Figure 5.2 above. Each name consists of a series of three-digit numbers representing the academic discipline, the journal, and the volume, followed by two two-digit numbers representing the issue and article. Thus in Figure 5.2 the highlighted file is the first article in the first issue of volume 723 of *Nuclear Physics A,* which is the fourth journal in the discipline of Nuclear Physics, which in turn is the ninth discipline in the corpus (collection was discontinued in some disciplines originally included), and so has the file name 0090047230101, being made up of 009, 004, 723, 0101.

5.3 Issues in determining corpus size

Measuring the precise size of a corpus containing data collected in the manner described above is not straightforward, and it should be stressed that the size of the corpus in this study cannot be stated with complete accuracy. There is something of a trade-off between number of tokens in an isolexical corpus and the accuracy with

which its size can be known. This section describes how this inaccuracy arises from both the software used to count the data and the methodology of collecting the data itself, and provides an estimate of how much influence each feature might have on the results of the study.

First, it is not certain how much of the text in a file containing a journal article actually should be considered as part of the article. Features such as the article title, abstract, numerical data in tables, the legends under figures and tables, bibliographical references, and footnotes and endnotes all contain text which, while to a greater or lesser extent related to the article, is not actually part of the writer's running prose. In some previous studies, e.g. Charles (2004), individual texts are "cleaned up" before inclusion in the corpus so that only running prose is counted. In the corpus in the present study, which contains thousands of files, manual "cleaning up" would be an extremely slow and lengthy process, and the longer it took, the less data could be collected, and the greater would be the risk of inconsistencies being introduced into the data as a result of human error. Since this study aims to investigate fixed collocational patterns rather than individual words, it is more appropriate to remove text which is not part of the running prose of an article later in the methodological process (see section 5.5 below). It was therefore decided to include in the corpus all text in each file.

5.3.1 Issues in determining corpus size: software

A basic source of uncertainty in calculating the size of the corpus is the software used. Instructing *Wordsmith Tools 4* or *Microsoft Word* to count the number of words in a file is a very simple operation, but no two applications ever seem to give the same result. Table 5.1 below shows word counts for a single text file by *Microsoft Word* and two different versions of *WordSmith Tools*. However, while it is clear that no two agree exactly, the difference between the largest and smallest counts is no more than half a percent.

Software	Word count
WS Tools 3	18,699
WS Tools 4	18,708
MS Word	18,614

Table 5.1: Word counts of the same text by different software

5.3.2 Issues in determining corpus size: symbol noise

In addition to the uncertainty that surrounds the amount of text which is part of a journal, and the fact that word counts by different software do not agree, there are other sources of uncertainty in the size of the corpus. One of these comes from the fact that computers represent characters as codes, in most cases the American Standard Code for Information Interchange (ASCII), based on the characters in the English alphabet. Academic articles often contain formulae and other characters which are not covered by the ASCII codes, and which when converted into plain text do not look like letters of the alphabet; these characters are here termed "symbol noise". Figure 5.3 below shows a text first saved in pdf format and also in html format, which has been converted into plain text format. It can clearly be seen that the two resulting plain text files are not identical:

original text in pdf:	converted to plain text:
The covariates vary over firms and (in some cases) over time for given firms: $X = (X_{11}, X_{12}, \ldots, X_{k1}, \ldots, X_{kI})'$ $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik_i})$ If we represent the disturbances as $u = (u_{11}, u_{12}, \ldots, u_{k_i I})$ then we can write the structural model to be estimated as $y = Xb + u.$	The covariates vary over firms and (in some cases) over time for given firms: □ ( ) X p x , x , … ,x , … ,x , 11 12 K1 KI ( ) x p x , x , … ,x . ki ki1 ki2 kini If we represent the disturbances as ( ) u p u ,u , … ,u , i i1 i2 ini then we can write the structural model to be estimated as y p Xb □ u.
original text in html:	**converted to plain text:**
The covariates vary over firms and (in some cases) over time for given firms: Multiple line equation(s) cannot be represented in ASCII text. If we represent the disturbances as Multiple line equation(s) cannot be represented in ASCII text. then we can write the structural model to be estimated as y = Xb + u.	The covariates vary over firms and (in some cases) over time for given firms: Multiple line equation(s) cannot be represented in ASCII text. If we represent the disturbances as Multiple line equation(s) cannot be represented in ASCII text. then we can write the structural model to be estimated as y = Xb + u.

Figure 5.3: Noise from non-ASCII characters in pdf and html files

The main distortion in word counts caused by symbol noise arises from "false positives": symbols being counted by the software as words. There are a number of ways in which *Wordsmith Tools* can be instructed by the user to identify a "word", but, for this study, a word is defined as a character or string of characters between two spaces. Symbol noise from pdf files is thus mistaken as words by the software, which inflates the reported size of the corpus. Text from journal articles originally in html format can contain a different form of symbol noise, as shown in the lower part of Figure 5.3. Since the special characters in formulae cannot be rendered in html, a nine-word standard phrase *Multiple line equation(s) cannot be represented in ASCII text* is used, which also inflates the word count. The nine words are included in the word count, even though the original formula may not be nine words long.

Single line equations which *can* be represented in plain text may still cause a minor distortion in the word count, as can be seen with the final equation in Figure 5.3 above. Conversion from pdf to plain text yields Example 5.1 below:

Example 5.1: y p Xb □ u.

Because it only contains four recognised characters, this is counted as four tokens by Wordsmith Tools. Conversion of the same equation from html to plain text, on the other hand, yields Example 5.2 below:

Example 5.2: y = Xb + u.

which the software counts as three tokens.

It is difficult to estimate how much of this type of symbol noise exists in the text files in the corpus. Such characters cannot be counted automatically by *Wordsmith Tools* because they cannot be recognised as search terms by the software. To find each occurrence of symbol noise, each text file would need to be manually read, a task which would be very time-consuming for a corpus of this size. In relation to the present study, it was decided to leave symbol noise in the texts uncorrected. Removing such noise may damage the integrity of the individual texts in the corpus because characters from formulae are often part of the syntax of a sentence, as in

the case of the = sign being used as the finite verb *equals*. They ought to be seen as part of the running prose of the text, and thus should indeed be counted as tokens, even if they are not traditionally viewed as "words."

Moreover, the potential of symbol noise to distort the results of the present study is reduced by the fact that this is a study of fixed collocational patterns rather than single words. Although symbol noise caused by this type of mathematical formula is likely to inflate reported totals of individual words, it is unlikely to have much effect on the number of fixed collocational patterns missed by the software because of the syntactic position of such formulae. These tend to occur at the end of clause or sentence elements, as in Example 5.3 below:

> Example 5.3: then we can write the structural model to be estimated as y = Xb + u.

5.3.3 Issues in determining corpus size: fluff

A more problematic type of symbol noise arises from the character codes used to represent paper texts electronically. When seriffed fonts such as Times New Roman are printed, some pairs of characters can be hard to read because the serifs in particular characters, such as the bulb on the end of the f character in Times New Roman, are very close to the adjacent character[2]. Examples of this in Times New Roman are ff, fi, and fl. Some computer character coding systems represent these pairs of characters, known as *digraphs*, as a single character, known as a *ligature*. In journal articles in two disciplines in this study, namely Mechanical Engineering and Philosophy, which both use the Unicode character encoding system, these digraphs are encoded using ligatures.

Since the ANSI (American National Standards Institute) character encoding used in the plain text files for this corpus does not have codes for ligatures, this meant that when articles in these disciplines are converted from pdf to plain text, the ligature character is replaced by the nearest possible substitute in the ANSI code. The Unicode ligature character representing *fi*, for example, is replaced by the ® symbol,

---

[2] This thesis is written in *Arial*, a "sans serif" font which is designed to be easy to read

and so the words *specific* and *configuration* would appear in text files as *speci®c* and *con®guration*. *WordSmith Tools* in turn does not recognise the ® character, and so counts *speci®c* as two tokens, *speci* and *_c*. This has the potential to inflate the word count.

This type of interference, termed here *fluff*, presents a more serious problem than symbol noise because it not only potentially inflates the reported word count, but also has the potential to affect the counting of fixed collocational patterns in the corpus. If the software is instructed to search for a particular pattern, such as *specific configuration*, which would be processed as a mixed string of characters and spaces, the fluff within the words in the pattern, in this case *speci®c con®guration*, will cause the software to miss the occurrence of the entire pattern, and thus it will be under-counted in the corpus.

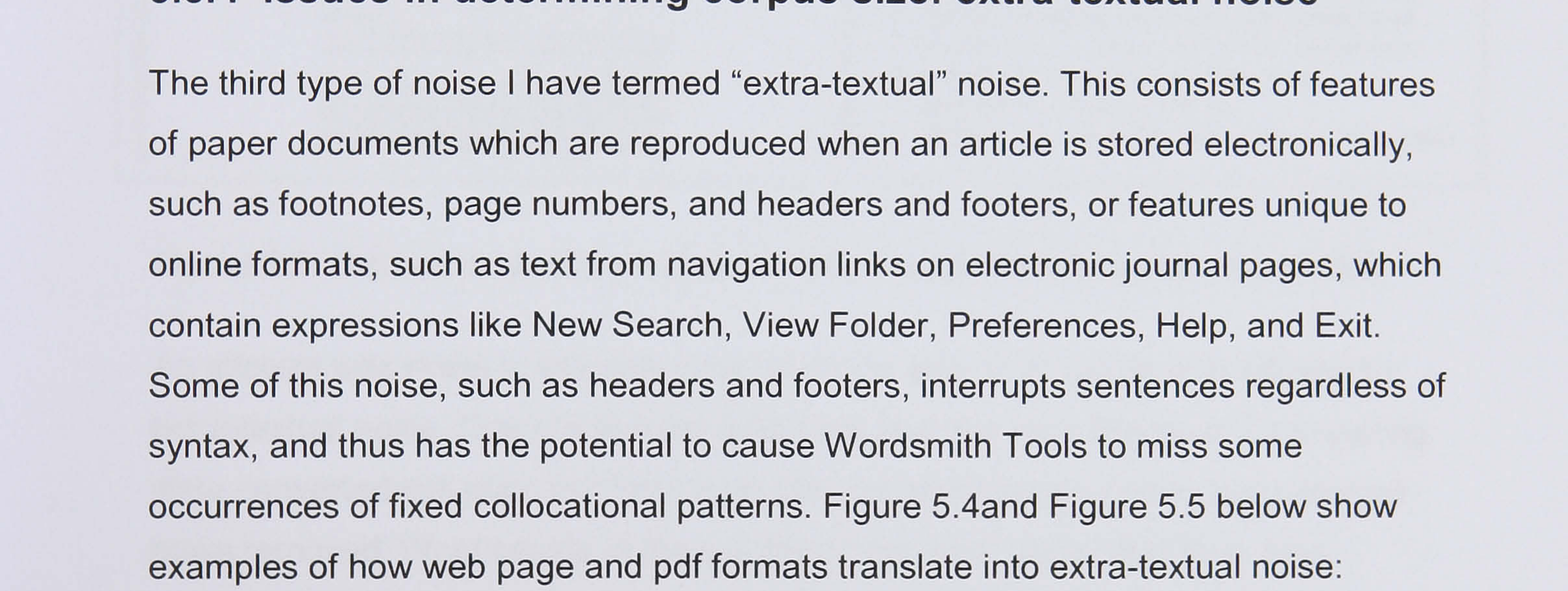
This problem could have been addressed by saving plain texts using the Unicode standard of electronic character encoding, rather than the older ANSI format. Unicode is a unique coding system which represents characters in such a way that they are recognised by any platform or software. Fluff would then have been recognised as the relevant ligature, rather than mis-stored as ® or _. When the corpus for the present study was being constructed, however, Unicode had not yet become standard in the software I was using: my version of *Notepad* only offered the option to save in ANSI format until 2004, when the option to save a file in Unicode format was also offered. *WordSmith Tools 4*, moreover, was not fully functional with Unicode until 2005, and so ANSI format was used instead. The use of ligatures rather than digraphs is not considered best practice in electronic text encoding precisely because it causes this kind of problem (Unicode 2008). Unicode has a code for *f* (U+0066) and a code for *l* (U+0006C) and so *fl* should be represented as a digraph (U+0066 + U+0006C) rather than as a ligature (U+FB02).

For the purposes of the present study, the influence of potential miscounts of such combinations resulting from fluff was judged to be acceptably low: *specific* occurs 14555 times, while *speci®c* and *speci? _c* occur a total of 751 times (5% of the total for *specific*); *configuration* occurs 5091 times while *con®guration* occurs 196 times (3.7% of the total for *specific*). Since this is a corpus-driven study which allows the

target items to emerge from the corpus, any frequent contiguous combinations which contain a word potentially affected by fluff, such as *the effect of*, will obviously be reported as occurring less frequently than actually is the case. However, fluff affects mainly two disciplines; in the quantitative analysis, the frequencies of occurrences of word combinations suspected of being affected by "fluff", and which are found to be frequent in the other six disciplines, will be more closely examined in Mechanical Engineering and Philosophy. The fact remains, however, that "fluff" has the potential to obscure some of the quantitative results of this study.

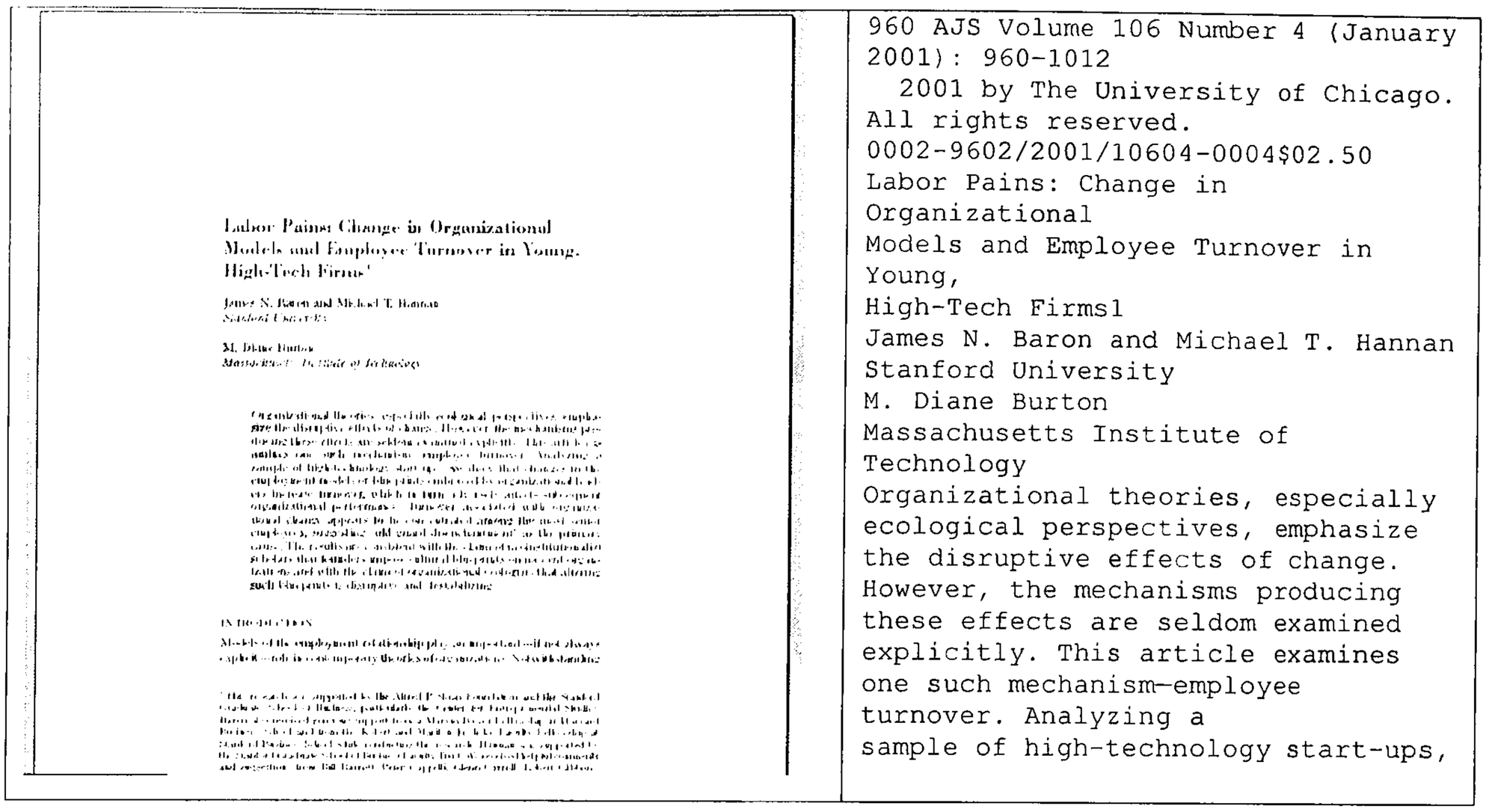
5.3.4 Issues in determining corpus size: extra-textual noise

The third type of noise I have termed "extra-textual" noise. This consists of features of paper documents which are reproduced when an article is stored electronically, such as footnotes, page numbers, and headers and footers, or features unique to online formats, such as text from navigation links on electronic journal pages, which contain expressions like New Search, View Folder, Preferences, Help, and Exit. Some of this noise, such as headers and footers, interrupts sentences regardless of syntax, and thus has the potential to cause Wordsmith Tools to miss some occurrences of fixed collocational patterns. Figure 5.4and Figure 5.5 below show examples of how web page and pdf formats translate into extra-textual noise:




```
New Search  | View Folder  |
Preferences  |  Help | Exit
 UNIVERSITY OF BIRMINGHAM


 Sign In to My EBSCOhost



  4 of 34    Result List | Refine Search
Print  E-mail  Save    Folder has items.

Formats:   Citation  HTML Full Text  PDF
Full Text  (229K)  Times Cited (1)
```

Figure 5.4: Extra-textual noise in plain text file caused by text from online navigation links in an html file



Labor Pains: Change in Organizational Models and Employee Turnover in Young, High-Tech Firms1

James N. Baron and Michael T. Hannan
Stanford University

M. Diane Burton
Massachusetts Institute of Technology

Organizational theories, especially ecological perspectives, emphasize the disruptive effects of change. However, the mechanisms producing these effects are seldom examined explicitly. This article examines one such mechanism—employee turnover. Analyzing a sample of high-technology start-ups,

960 AJS Volume 106 Number 4 (January 2001): 960-1012
2001 by The University of Chicago. All rights reserved.
0002-9602/2001/10604-0004$02.50

Figure 5.5: Extra-textual noise in plain text file caused by the header in a pdf file

An attempt was made to estimate how much the size of an article was inflated by extra-textual noise. One article from Sociology and one from Electrical Engineering were converted into plain text from both pdf and html formats and all extra-textual noise removed. Word counts of the resulting "raw" and "clean" text files were performed by three different software applications. The amount of noise in the Sociology article was found to be around 13%, while the amount of noise in the Electrical Engineering article was around 5%, as in Table 5.2 below:

Software	raw pdf	clean pdf	"noise" in pdf (%)		raw html	clean html	"noise" in html (%)
Sociology article							
WS Tools 3	21,617	18,780	13.1		21,647	18,699	13.6
WS Tools 4	21,617	18,780	13.1		21,685	18,708	13.7
MS Word	24,029	18,592	22.6		21,415	18,614	13.1
Electrical Engineering article							
WS Tools 3	2,535	2,405	5.1		2,602	2,454	5.7
WS Tools 4	2,536	2,405	5.2		2,602	2,454	5.7
MS Word	2,390	2,298	3.8		2,548	2,390	6.2

Table 5.2: Effects of extra-textual noise on calculations of Sociology and Electrical Engineering article size

While the amount of noise is different in these disciplines, it is estimated that the effects of extra-textual noise in different subcorpora will not be sufficient to unbalance the different sizes. Articles in Sociology journals are much longer than texts in Electrical Engineering journals, and fewer texts are required to collect a million words of Sociology than are required to collect a million words of Electrical Engineering. As each article page has a header and a footer, a thirty-page Sociology article has 60 headers and footers while a four-page Electrical Engineering text has eight. One hundred Sociology texts and 750 Electrical Engineering texts would therefore both contain 6000 headers and footers. It was decided to over-collect texts in each journal by around five percent, so that a subcorpus of 1,050,000 tokens from five journals would count as 5,250,000 tokens, which will be taken to represent around five million running words without the noise.

In three disciplines, Electrical Engineering, Nuclear Physics, and Philosophy, the majority of texts in the subcorpus were collected from articles published in pdf format rather than html, and as a result these discipline subcorpora are likely to contain more extra-textual noise than in subcorpora from disciplines whose texts were collected from articles published in html format. To allow for this greater amount of noise around 20 percent more data was accordingly collected from each of the five journals texts in Electrical Engineering and Nuclear Physics, and so the token count taken to represent five million running words for these two disciplines is nearer six million. Philosophy journal articles were seen to be different in character from those in Electrical Engineering and Nuclear Physics, as Philosophy articles contain much more running prose and almost no tables and diagrams and thus less extra-textual noise. As in the case of Sociology and Electrical Engineering, it was decided to overcollect tokens from each Philosophy journal by 5 percent, so that a subcorpus of 1,050,000 tokens from five journals would total as 5,250,000 tokens, which will be taken to represent around five million running words without the noise.

A summary of the composition of the research article corpus which was collected according to the procedure outlined above is shown in Appendix 2. The token counts for both versions of each subcorpus, including the original format of the article from which the text was collected, is shown in Appendix 3.

The research article corpus has achieved its aim of being several times larger than corpora collected for previous studies of academic English. Figure 5.6 on the next page shows a comparison between the isolexical version of the research article corpus and the "academic" parts of previous off-the-shelf corpora discussed in section 4.3 above. In addition to its larger size, the research article corpus also provides a more equal balance between different disciplines, and, allowing for different sources of noise in texts in each discipline, there are an equal number of tokens in each discipline. Only in the US textbooks section of the Bank of English does any discipline contain as many tokens as one journal in the research article corpus. Figure 5.6 also compares the research article corpus with the section of the T2K-SWAL corpus in which Biber (2006) identified the functions of lexical bundles in the genre of the academic textbook, and also Hyland's academic discourse corpus (Hyland 2008) in which his functional framework for lexical bundles was developed. It can be seen that only the Applied Linguistics section of Hyland's corpus contains as many tokens as the journals in the research article corpus.

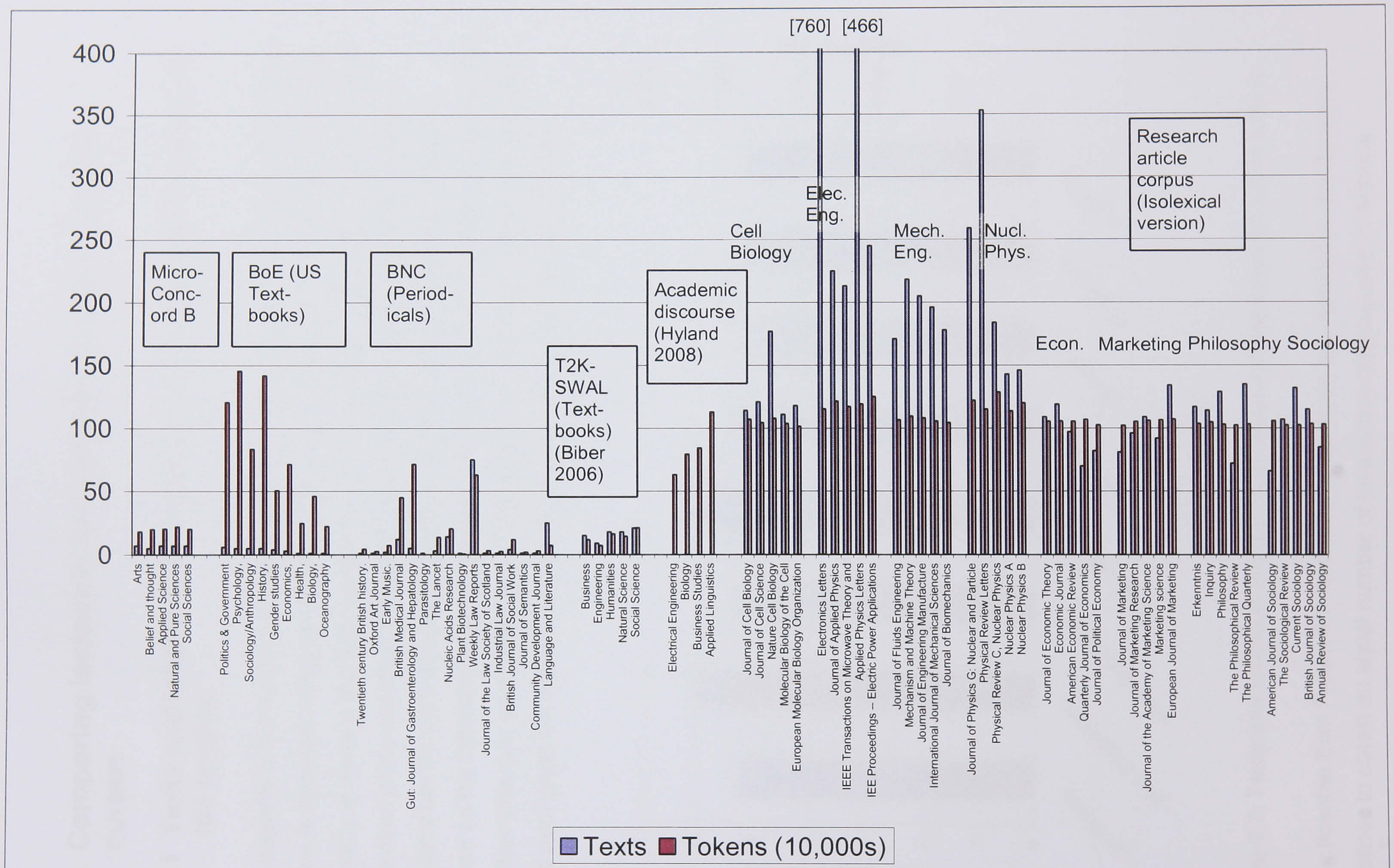


Figure 5.6: The isolexical version of the research article corpus compared with other academic English corpora

5.4 Comparing isolexical and isotextual versions of the corpus

5.4.1 Texts and tokens in the isolexical and isotextual versions of the corpus

The disparity in the number of texts in each subcorpus in the isolexical version of the corpus is apparent in Figure 5.6, (two journals have 760 and 466 texts each which is too many to show in the figure) but is shown more clearly in Figure 5.7 below. This is a bar chart of the isolexical version of the corpus which shows texts and tokens in the same bar for each discipline. The blue bar represents the number of tokens (shown on the left-hand axis); these totals include noise and represent approximately 5 million tokens in each subcorpus. The red portion of each bar (shown on the right-hand axis) gives the number of texts required to collect this number of tokens.

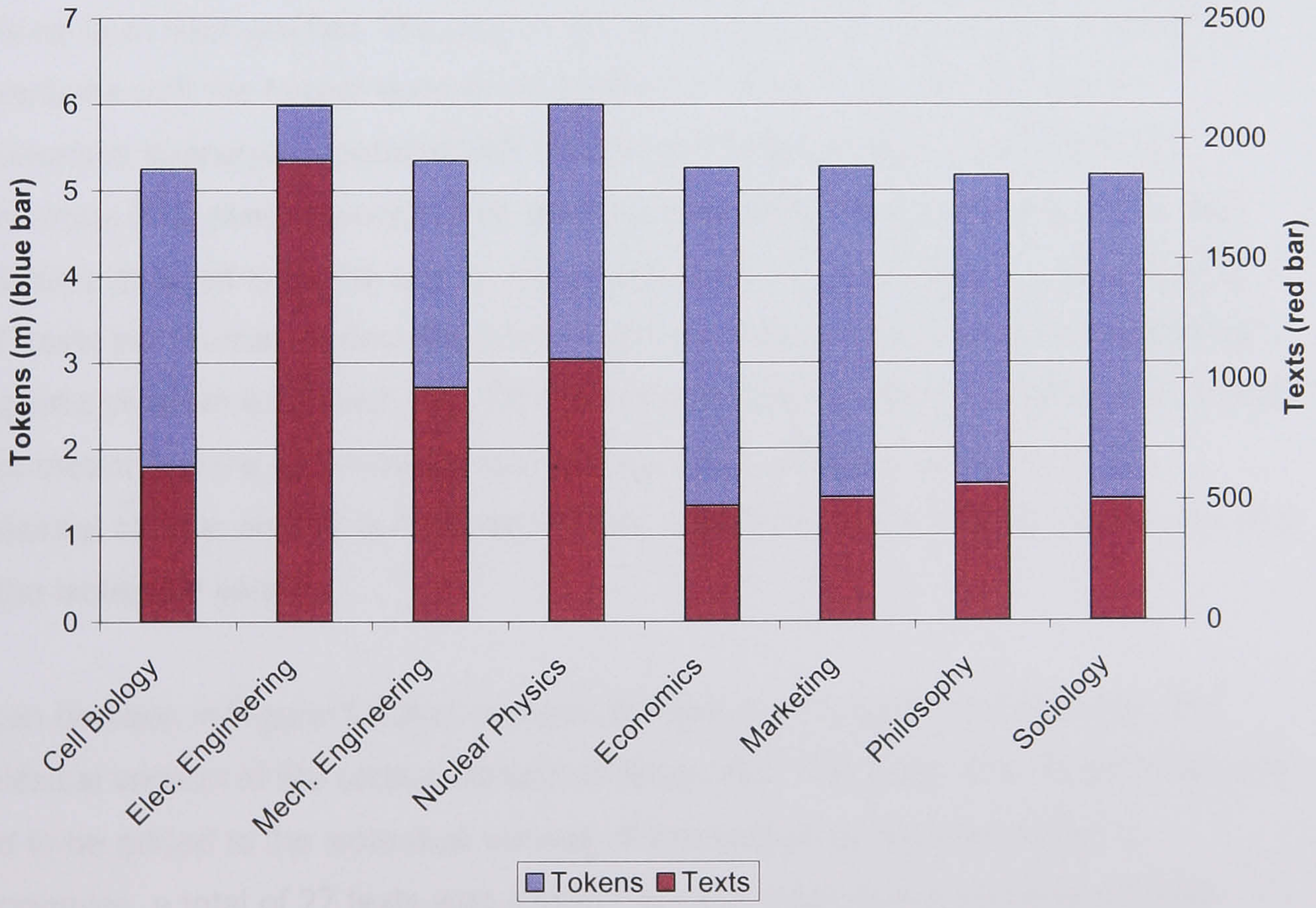


Figure 5.7: Texts and tokens in the isolexical version of the research article corpus

It is noticeable from Figure 5.7 that many fewer texts are required in non-Science disciplines to obtain an equal number of tokens than are required in Science

disciplines. It is this difference that may bias results of comparisons between isolexical subcorpora between disciplines. If a linguistic feature is found to be more common in the Electrical Engineering isolexical subcorpus than in subcorpora from other disciplines, the finding may not be due to the nature of the language of Electrical Engineering *per se*, i.e. something occurs more often in a given amount of language of Electrical Engineering than it does in a given amount of language from another discipline, but may instead be due to the fact that the Electrical Engineering isolexical subcorpus contains more than twice as many texts as the subcorpora for the other disciplines. Comparing such isolexical results with comparisons between isotextual subcorpora from different disciplines will make it possible to observe whether this imbalance in the number of texts influences the results.

To allow isolexical results to be compared with isotextual results, an isotextual version of the research article corpus was prepared. This in effect meant choosing a subset of the isolexical version for each discipline which comprised an equal number of texts from each journal. The size of the isotextual subset was determined by the disciplines with the fewest texts in the isolexical subcorpora. The isolexical Economics subcorpus contained 477 texts (see Appendix 3.5), the Marketing subcorpus 512 (see Appendix 3.6), and Sociology 505 (see Appendix 3.8). It was therefore decided to fix the size of each isotextual corpus at 500 texts, comprising 100 texts per journal. In deciding which texts to include, a chronological rather than linguistic criterion was used: the 100 most recent texts in each journal were included. This meant that the isolexical version of the corpus is largely a subset of the isolexical corpus, so that in most disciplines, texts from the isolexical version are also in the isotextual version.

It can be seen in Figure 5.6 that in some journals in non-Science disciplines, the isolexical version of the corpus contained fewer than 100 texts. This meant that texts had to be added to the isotextual version of the corpus for that discipline. In Economics, a total of 27 texts was added from the most recent volumes of three Economics journals to bring the total number of texts for those journals up to 100 (see Appendix 3.5). In Marketing, 31 texts were added in total, although 12 were also removed from other journals (see Appendix 3.6) to bring the total number of texts for that journal down to 100. In Philosophy, 18 texts were added for one journal, while

67 were taken away from other journals (see Appendix 3.7). In Sociology, 51 texts were added from the most recent volumes of two journals to bring the total number of texts for those journals up to 100 while 5 were taken away from other journals (see Appendix 3.8).

Figure 5.8 below is a bar chart of the isotextual version of the corpus which shows texts and tokens in the same bar for each discipline. The red bar represents the number of texts (shown on the left-hand axis); The blue portion of each bar (shown on the right-hand axis) gives the number of tokens collected by including 500 texts. The token totals again include noise.

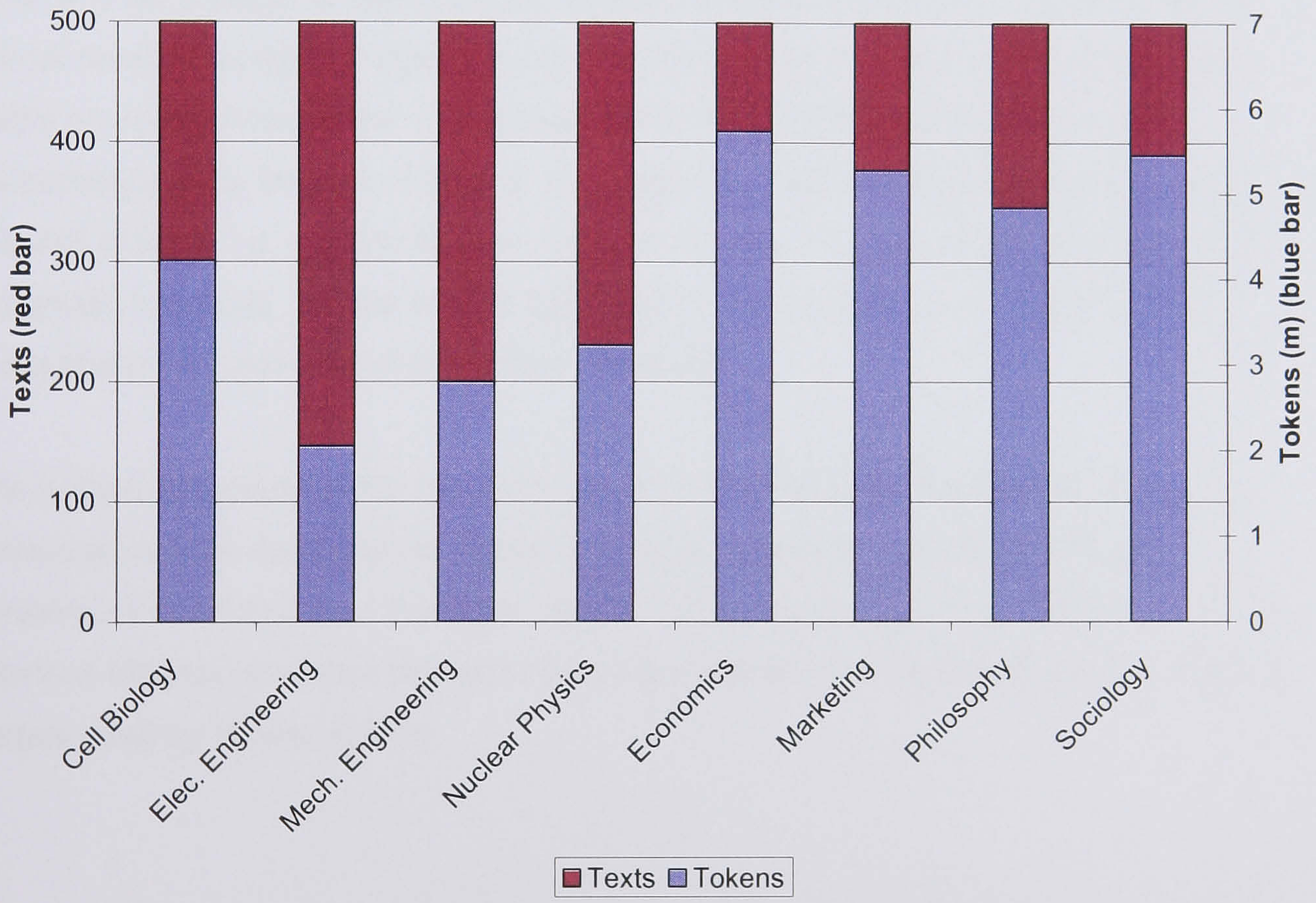


Figure 5.8: Texts and tokens in the isotextual version of the research article corpus

It can be seen from Figure 5.8 that, as expected, collecting a similar number of texts in each subcorpus results in an imbalance between the number of tokens in each discipline. 500 texts in non-Science disciplines contain more tokens than 500 texts in Science disciplines. This difference may again bias results of comparisons between isotextual subcorpora between disciplines. If a linguistic feature is found to be less

common in the Electrical Engineering isotextual subcorpus than in subcorpora from other disciplines, for example, the finding again may not be due to the language of Electrical Engineering *per se*, but may instead be due to the fact that the Electrical Engineering isotextual subcorpus contains many fewer tokens than the subcorpora for the other disciplines. Comparing such results with comparisons between isotextual subcorpora of different disciplines will make it possible to observe whether the disparity between the number of tokens has affected the result.

In this thesis, findings from the isotextual version of the corpus will be viewed as being more revealing than findings from the isolexical version. This is because a text is a communicative event, written in order to achieve a pragmatic, communicative goal, and will contain all the functions needed to achieve that goal. The literature on lexical bundles assigns these fixed collocational patterns a pragmatic function, yet relies on corpora measured in language rather than communication. A corpus containing a large amount of tokens, i.e. language, and a corpus containing a large number of texts, i.e. communication, will both contain a large number of instances of pragmatic functions, but the corpus balanced for texts should, it is hoped, reveal more about the instances of pragmatic functions.

The isotextual version of the research article corpus contains fewer tokens than the isolexical version. As shown in Figure 5.9 below, however, it nonetheless still contains more tokens than the other corpora of academic discourse mentioned in the previous chapter, and also contains more tokens than the T2K-SWAL corpus and the corpus used by Hyland (2008).

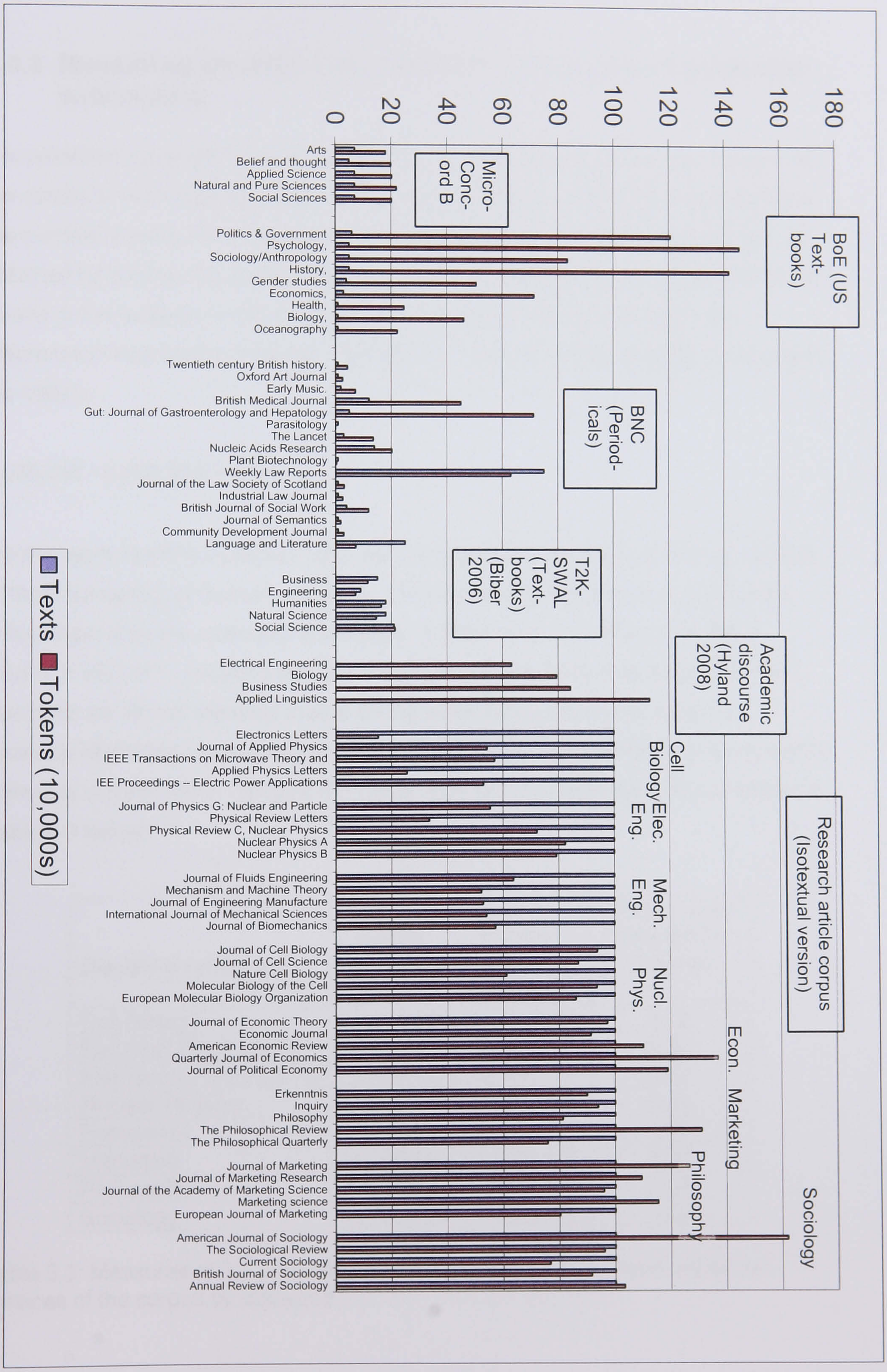


Figure 5.9: The isotextual version of the research article corpus compared with other academic English corpora

5.4.2 Measuring the difference between isolexical and isotextual subcorpora

It is possible to quantify the relative size of the isolexical and isotextual versions of the corpus in each discipline, both in terms of the number of tokens and in terms of the number of texts. This measure, here called a version size differential (VSD), is obtained by dividing the number of tokens in the isotextual corpus by the number of tokens in the isolexical corpus. The closer the VSD is to zero, the larger the difference in size between the two versions. For Electrical Engineering, for example, the VSD is

2,050,397 / 5,983,954 = 0.34

which means that the isotextual Engineering subcorpus is around a third of the size of the Engineering isolexical corpus, i.e. the isolexical subcorpus is around three times larger than the isotextual subcorpus. In Marketing, the same calculation returns a VSD of 1, meaning that the isolexical and isotextual subcorpora for this discipline are almost identical in size on the other hand. This is because the isolexical Marketing subcorpus contained 505 texts, which is almost the same as the isotextual version which contains 500 texts. The VSD for each discipline is shown in Table 5.3 below:

Discipline subcorpus	Tokens (isolexical version)	Tokens (isotextual version)	Version size differential (tokens)
Cell Biology	5,248,895	4,210,833	0.80
Electrical Engineering	5,983,954	2,050,397	0.34
Mechanical Engineering	5,345,399	2,798,202	0.52
Nuclear Physics	5,999,688	3,225,066	0.54
Economics	5,261,076	5,728,437	1.09
Marketing	5,274,564	5,275,402	1.00
Philosophy	5,173,961	4,840,725	0.94
Sociology	5,181,285	5,453,244	1.05

Table 5.3: Measures of the difference in size between isolexical and isotextual versions of the corpus by discipline

The VSD measure reflects the difference in size between the two versions, and will allow any observed differences in the frequencies of fixed collocational patterns to be checked against the differences in size between the two subcorpora. It would of course be possible to measure the difference between the two versions in terms of the number of texts in each subcorpus, but since this study is starting with the isolexical and comparing with the isotextual, it is felt that the VSD in terms of tokens is more useful.

The next section of this chapter describes the procedure used for identifying fixed collocational patterns in both versions of the research article corpus.

5.5 Identification procedure for fixed collocational patterns

5.5.1 The *WordSmith Tools* cluster function

As discussed in section 4.2, this study follows a corpus-driven methodology: the items which are under investigation will emerge from the data without prior annotation of the data by the researcher. It will be recalled that the items required to answer the first research question are the most frequently occurring fixed collocational patterns in research articles across all eight disciplines. In order to identify these targets, it was first necessary to identify the most frequent fixed collocational patterns in each discipline, then find which of these occur in all eight disciplines.

Once the two versions of the corpus had been collected, a process which took just over two years, an index was created for the subcorpus for each discipline. An index consists of two files which in effect contain all the information from the subcorpus in a reduced form. One file contains a list of word types found to occur in the subcorpus, while the other contains information about the position of each token in the subcorpus. The index files contain less information than the text files in the subcorpus, and are therefore smaller, and so can be searched more quickly by the lexical analysis software than the whole subcorpus could be. Indexes were created for each discipline subcorpus in both versions of the corpus.

Once the index files were created, the *Wordsmith Tools* cluster function was used in order to obtain frequencies of fixed collocational patterns in each subcorpus index. It is important to be precise about terminology: the term *cluster* is used in this section since it is the theory-neutral name used in *WordSmith Tools* to denote a string of spaces and characters with a space either side. It will be decided later in the study whether the clusters found using *Wordsmith Tools* can be judged to be fixed collocational patterns or not.

The cluster search can include or exclude certain clusters depending on the search criteria used. These criteria can be set by the user and relate to whether clusters including punctuation are counted, whether strings including hyphens are regarded as words, and how often a cluster has to occur before it is included in the frequency list. The software was instructed to ignore punctuation, so that clusters that included a sentence boundary would still be counted. This meant that possible n-grams, which by definition can cross sentence or turn boundaries, would be counted as well as possible lexical bundles, which cannot cross sentence boundaries. The inclusion of clusters which contain punctuation also helps prevent accidental undercounting, such as when the software misinterprets a decimal point as a full stop and excludes a cluster from the frequency count.

The software also allows the user to set a maximum and minimum number of occurrences of clusters; no maximum limit was set, but a minimum cut-off point of 100 occurrences in each discipline was specified. In the isolexical version of the corpus, this equates to 20 occurrences per million words, which is between the cut-off point of 10 per million words in Biber et al. (1999: 992) and the cut-off point of 40 per million words of Biber et al. (2004: 376). In the isotextual version this cut-off point of 100 occurrences per discipline equates to 20 occurrences per 100 texts. Also following Biber, the software was instructed to ignore clusters which occurred in fewer than 5 texts, in order to avoid counting idiosyncratic uses by a small number of authors.

Using the cluster function, frequency tables were obtained for 3, 4, 5, 6, 7, and 8-word clusters in each subcorpus in both versions of the corpus. These frequency tables included clusters which were part of extra-textual noise (see 5.3.1 above).

Clusters appearing to contain noise were checked and then, if confirmed to be noisy, were discarded. Table 5.4 below illustrates how this is done with an example of results returned by the *Wordsmith Tools* cluster function for 4-word clusters, in this case for Philosophy and Nuclear Physics:

	Philosophy	**freq.**	**Nuclear Physics**	**freq.**
1	*the editors of the*	1484	*et al phys rev*	2152
2	*editors of the philosophical*	1254	as a function of	2092
3	in the case of	826	*et al nuclear physics*	1975
4	on the other hand	730	*al nuclear physics a*	1867
5	on the basis of	701	*summaryplus full text links*	1189
6	that there is a	678	in the case of	1182
7	*of the philosophical quarterly*	656	*full text via crossref*	1111
8	if and only if	545	*al phys rev c*	1015
9	that there is no	499	*full text links pdf*	994
10	in terms of the	453	on the other hand	983
11	at the same time	427	*al phys rev lett*	954
12	in the sense that	385	a function of the	869
13	*cambridge cambridge university press*	368	in terms of the	860
14	the fact that the	362	*et al nucl phys*	810
15	on the one hand	352	with respect to the	757
16	it is true that	339	*j phys g nucl*	634
17	in such a way	311	*et al phys lett*	607
18	*oxford oxford university press*	310	*g nucl part phys*	584
19	that it is not	302	*phys g nucl part*	584
20	in so far as	295	*the american physical society*	537

Table 5.4: Noise in frequency lists for 4-word isotextual clusters in Philosophy and Nuclear Physics (shown in italics)

It can be seen from the table that several of the top 20 4-word clusters appear to be noise, rather than part of the running text. The methodological advantages of leaving this type of noise in the corpus, rather than deleting it manually file by file before any searches are performed, can now be seen. In most cases it is clear which clusters are noise and which are part of running text: *cambridge cambridge university press* and *et al phys lett* are both part of a bibliographical reference rather than running text, while *summaryplus full text links* comes from a sequence of navigational links in an article in html format. All of these examples can be discarded from the frequency list, which means being able to discard several hundred instances of noise all at once.

In cases where it is not immediately clear whether the combination is noise or part of the running text of the article, e.g. *the editors of the*, there are two ways of determining whether it should be discarded. The first is the "plot" function in WordSmith Tools, which displays a graphical representation, from left to right, of the occurrences of a feature in each text in the corpus. Figure 5.10 shows how occurrences of *the editor of the* are evenly spaced in each text. It is in fact part of a standard copyright notice which is repeated at the foot of every page in all articles and which thus counts as extra-textual noise rather than running prose.

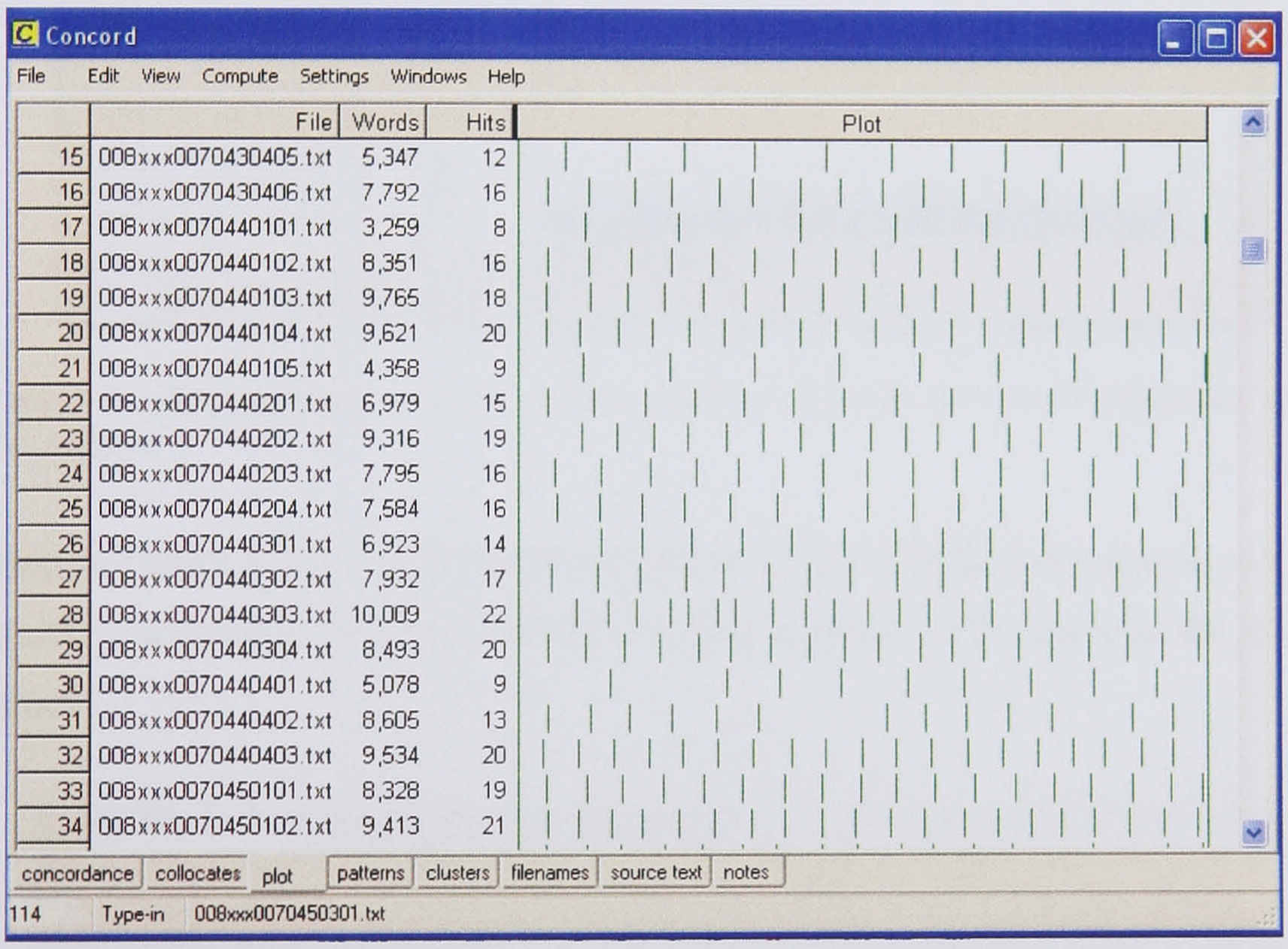


Figure 5.10: Regularity of noisy clusters in Philosophy articles: distribution of *the editors of the* as shown by *WordSmith Tools* plot function

The regularity of occurrence displayed in Figure 5.10 can be contrasted with the plot for a non-noisy cluster *in the case of the* displayed in Figure 5.11 below. The location of this pattern in the various texts is less regular, and so quickly shows that this cluster is not noise.

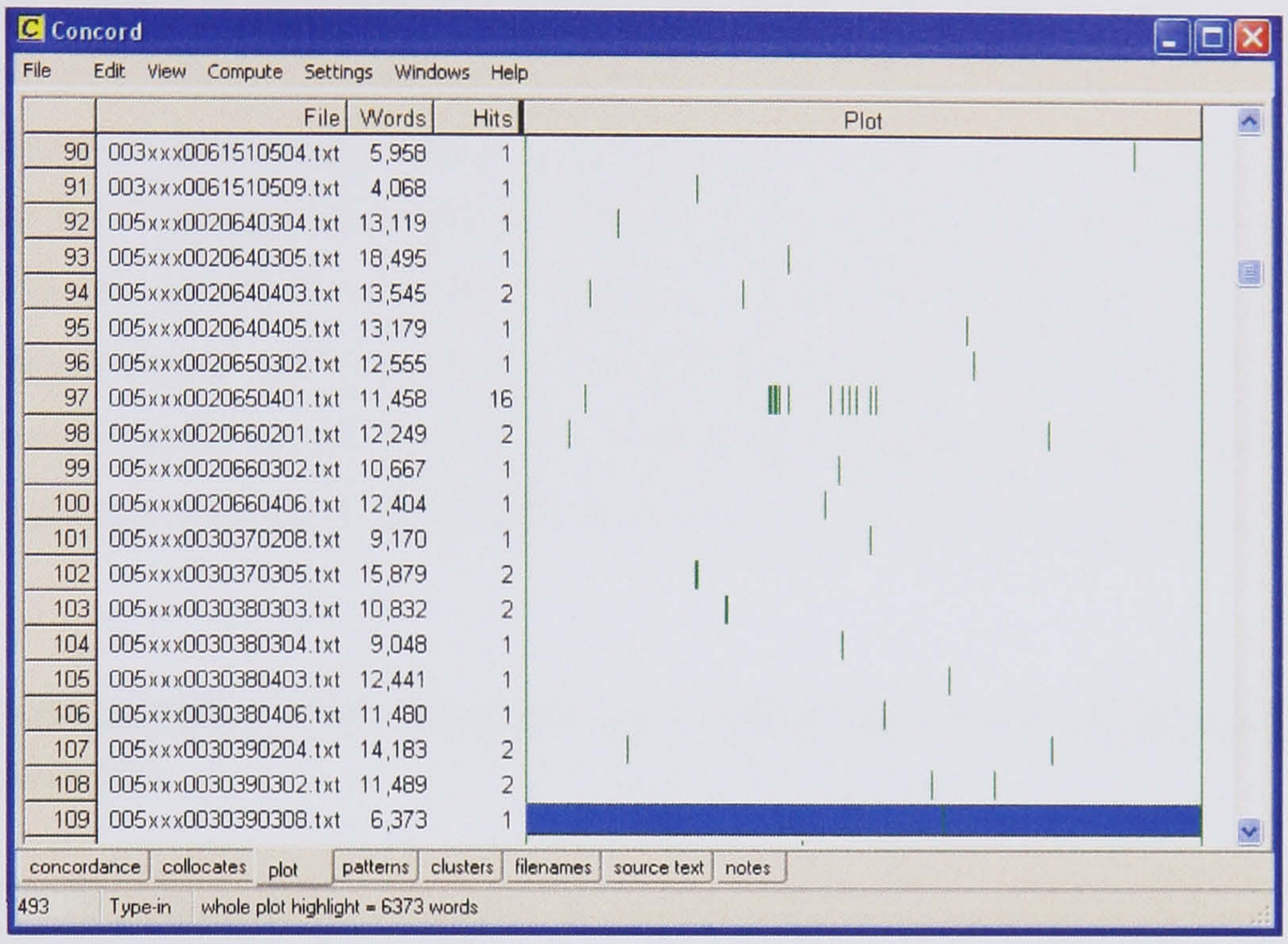


Figure 5.11: Distribution of *in the case of the* as shown by the WordSmith Tools plot function

The area in Figure 5.11 where the plot lines are close together represents the paragraph in Example 5.4 from a Marketing text in which *in the case of the* is used very often.

> Example 5.4: In the case of the brand Alpha, displays and advertising information were not included, because these factors were active in less than 1% of the purchases made in this category. In the case of the brand Zeta, the display information is not included for the same reason. The reciprocal effect variable (IND) is significant at the 1% risk level for extensions Alpha and Zeta, but it is not significant for brand Gamma. The odds ratio for the reciprocal effect variable is 16 in the case of the brand Alpha and 52 in the case of the brand Zeta.(10) This suggests that in the case of the brand Alpha, extension trial enhances the odds of purchasing the parent brand as opposed to some other brand by 16 times. Similarly, in the case of the brand Zeta, extension trial enhances the odds of purchasing the parent brand by 52 times.
>
> (Marketing 0050020650401)

If the plot function did not confirm whether a cluster was noise, then the second way of determining whether or not it was noisy was to produce a concordance of the cluster using the concordance tool. This allowed it to be seen in context and its statues determined. If the cluster was found to be noisy it was deleted from the frequency list.

Deleting noise from frequency lists at this point in the study meant that it was possible to discard hundreds of occurrences of false positives, i.e. clusters arising from noise, all at once. This way of discarding noise is much more time-efficient than individually pruning copyright notices, headers and footers and so on from each of the thousands of articles in the corpus. This method is also more consistent and reliable: the amount of time required to clean up all individual texts would have undoubtedly allowed human error into the pruning process. This noise-pruning procedure was performed for 3, 4, 5, 6, 7, and 8-word clusters in each of the eight academic disciplines in both the isolexical and isotextual versions of the corpus.

5.5.2 Quantitative identification procedure for fixed collocational patterns

Once the noisy clusters were removed from the WordSmith cluster frequency tables, the tables were saved as Microsoft Excel spreadsheets. Spreadsheets are useful for performing simple statistical functions on frequency data, and can produce charts to illustrate distributions graphically. Excel was most useful for identifying fixed collocational patterns which occurred across all 8 disciplines in both versions of the corpus. In the case of 3-word patterns, for example, this was done by inserting the 3-word cluster frequency lists for all 8 disciplines in one version into a single column in a Microsoft Excel spreadsheet, in this case column A, resulting in a spreadsheet with nearly 19,000 rows. The cells in the adjacent rows all contained a conditional formula (COUNTIF(A:A,A2)=8). The conditional formula counts the number of occurrences of the patterns in column A and, if the total number of occurrences of a pattern in column A is equal to 8, changes the value of each cell adjacent to that pattern to “true”. If the number of occurrences of a pattern is less than 8, the value of the adjacent cell is “false”. It was possible to extract the combinations adjacent to a “true” result by use of a sort command, and thus identify those fixed collocational patterns which occur more than 100 times in all 8 disciplines. Finally, these combinations were ranked in order of their frequency in the corpus as a whole.

5.5.3 The statistical significance of fixed collocational patterns

It was stated in section 1.3 that absolute frequency of occurrence (Kjellmer 1984) will be regarded as an indication that a fixed collocational pattern is important. Many studies of collocations, however, use a variety of statistical measures to evaluate the probability that a particular combination of words is not only important but also significant, i.e. a statistically calculated statement that the collocation is extremely unlikely to occur by chance and is therefore not a random occurrence. This can be determined by comparing the observed frequency of occurrence of a collocation (how often one of the words occurs with the other word), with its expected frequency of occurrence (how often each word occurs in the corpus). The most commonly used measurements of collocational significance are the Mutual Information score (Church & Hanks 1990) which indicates the strength of a collocational relationship between two words. For the present study, however, it was assumed that the size of the corpus was sufficient for absolute frequencies of occurrence of important fixed collocational patterns to be noticeably greater than for less important patterns.

The assumptions behind these statistical methods, moreover, are open to criticism. Statistical approaches to language treat words in a similar way: a collection of texts is not seen as a purposeful communicative act, but as an abstract set of data containing information about a series of events which have occurred. When a word is used in a text its statistical equivalent is the occurrence of an independent event with a probability attached to it, such as tossing a coin (Stubbs 2001: 73). Statistical methods test the strength of relationships between what are assumed to be independent events in the data set, and then these are re-expressed as generalisations about the language. The test for significance is that a collocation occurs much more often than would "otherwise be expected." A statistically significant collocation is extremely unlikely to occur by chance and is therefore not a random occurrence. However words do not occur at random in a corpus, and they are not independent events. Language data is a record of the instantiation of goal-oriented communicative acts, and it is difficult to imagine what might be the "expected" number of occurrences of a word or fixed collocational pattern in the first place.

It was thought therefore, that calculating the statistical significance of fixed collocational patterns in this study would not be useful. The larger size of this corpus means that the numbers of occurrences are consequently larger than in previous studies, meaning that the frequencies of fixed collocational patterns stand as they are, and do not need to be normed or otherwise processed. This lends the results of this study a robustness not found in other studies with normed frequencies.

5.5.4 Qualitative identification procedure for fixed collocational patterns

The third research question of this thesis involves investigating in more detail the discourse function of frequent fixed collocational patterns. While Corpus Linguistics methodology enabled these combinations to be identified quantitatively across disciplines, the qualitative investigation of the use of these patterns in the discourse must be analysed manually. Software can be instructed to count but cannot analyse for semantics and pragmatics no matter how much context is available. The methodological approach developed in this section thus moves from corpus-driven linguistics to a more corpus-based text and discourse analysis.

Concordance lines as seen above can reveal more about the meaning or use of a word from a screen's width containing around 200 characters, but such concordance lines do not provide enough context for the textual aspects of the use of each occurrence to be discerned. The "grow" function of the concordancer in Wordsmith Tools can only increase the length of a concordance line to a maximum of 2000 characters, which equates to around a paragraph of surrounding text.

There are two ways of determining the textual location of a pattern. First the "source text" view of the Wordsmith concordancing tool allows any plain text file from the corpus to be viewed in its entirety. Figure 5.12 below shows how, using the "source text" function, an occurrence of *in the case of the* was identified as being near the Introduction section of an Electrical Engineering article. The other way is to look in the text itself, either in plain text format (for files only available in html format) or in pdf format. In the case of files displayed in pdf format, this is often a more revealing way to examine the context of each pattern, since the pdf format preserves

typographical features of the text such as layout, formatting, and equations. As we have seen, in plain text files these can be displayed as noise and can obscure the relevant parts of the text to the analyst.

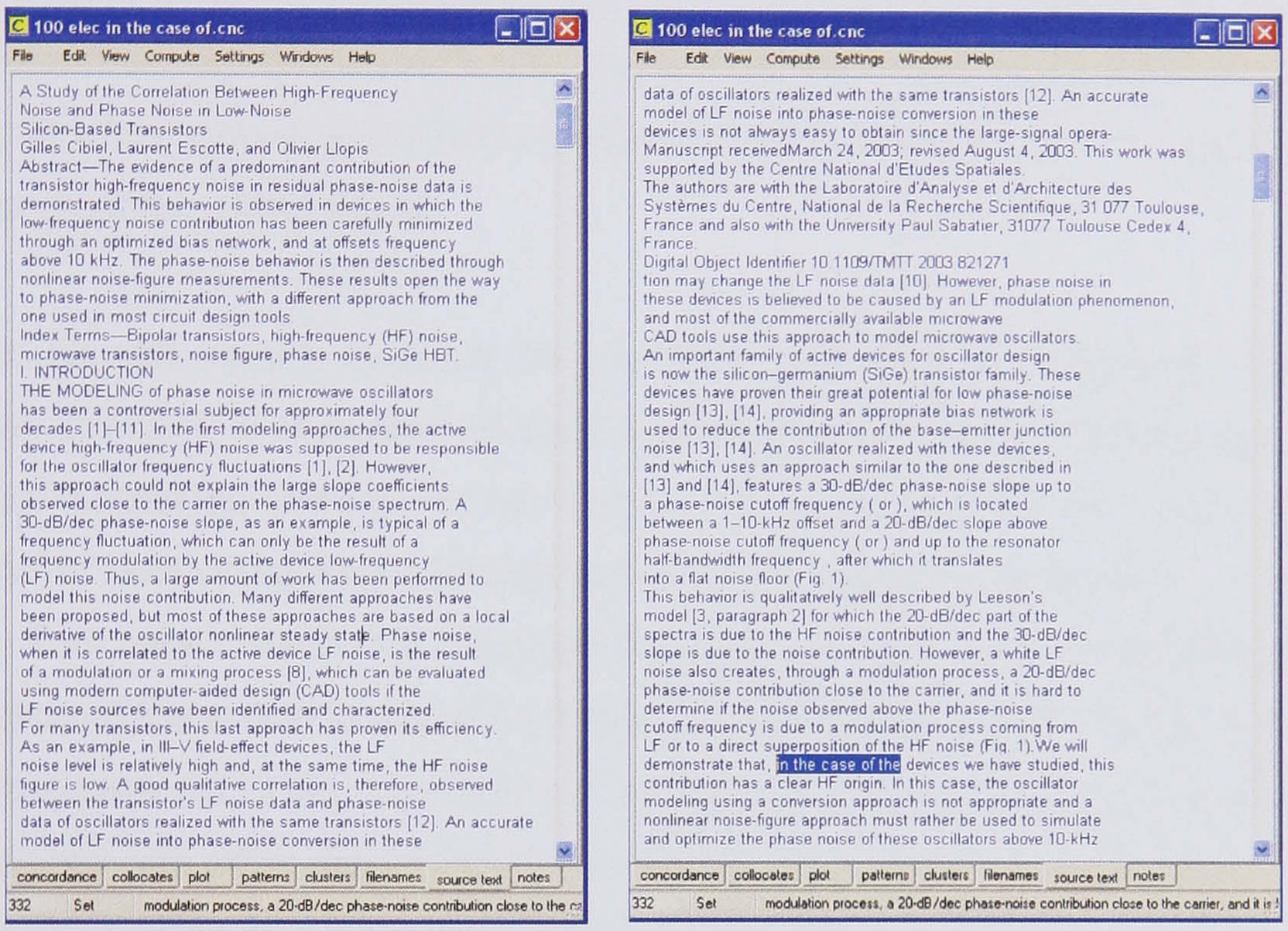


Figure 5.12: Identification of the textual location of *in the case of the* using the "source text" view in WordSmith Concord

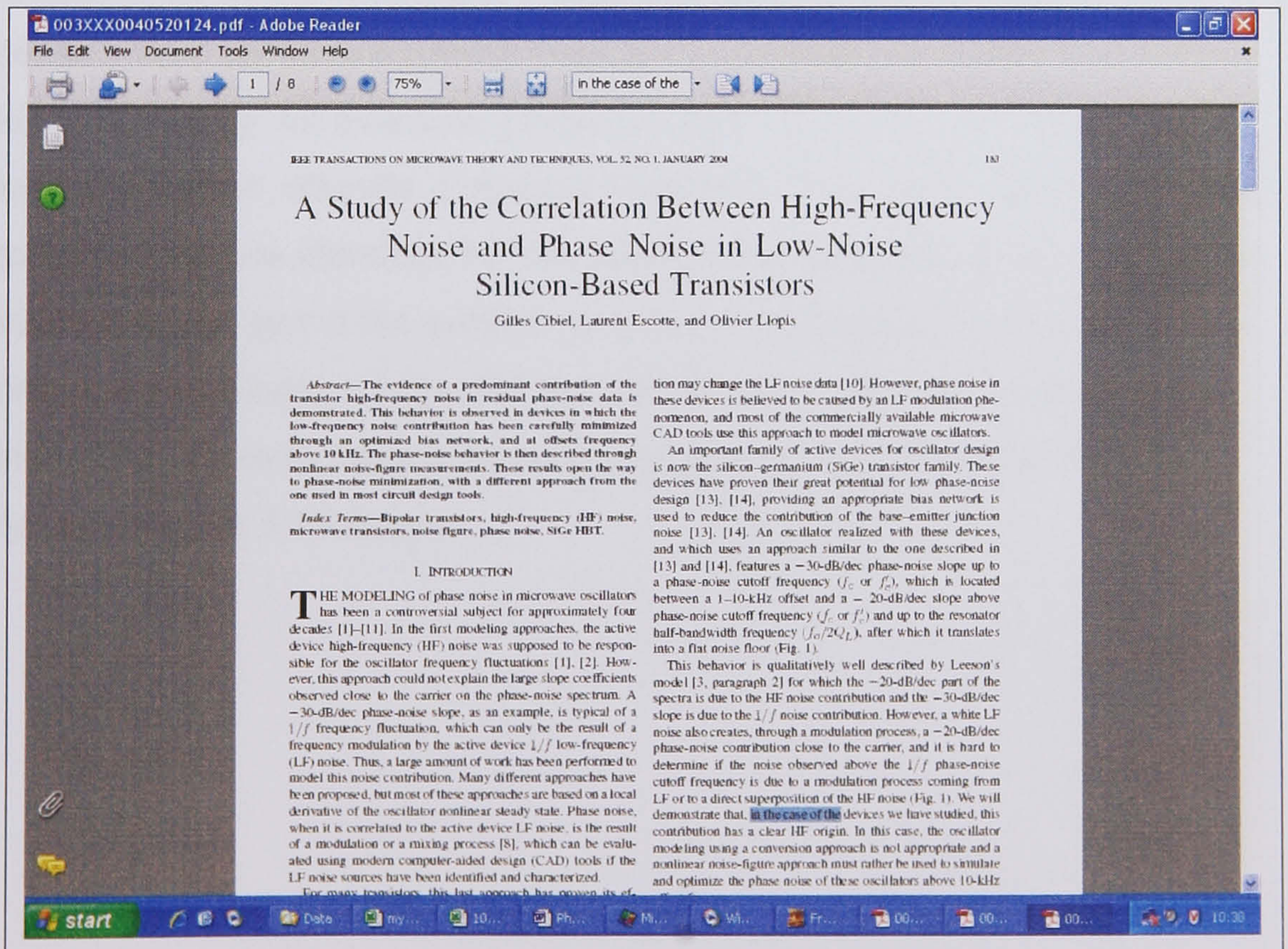
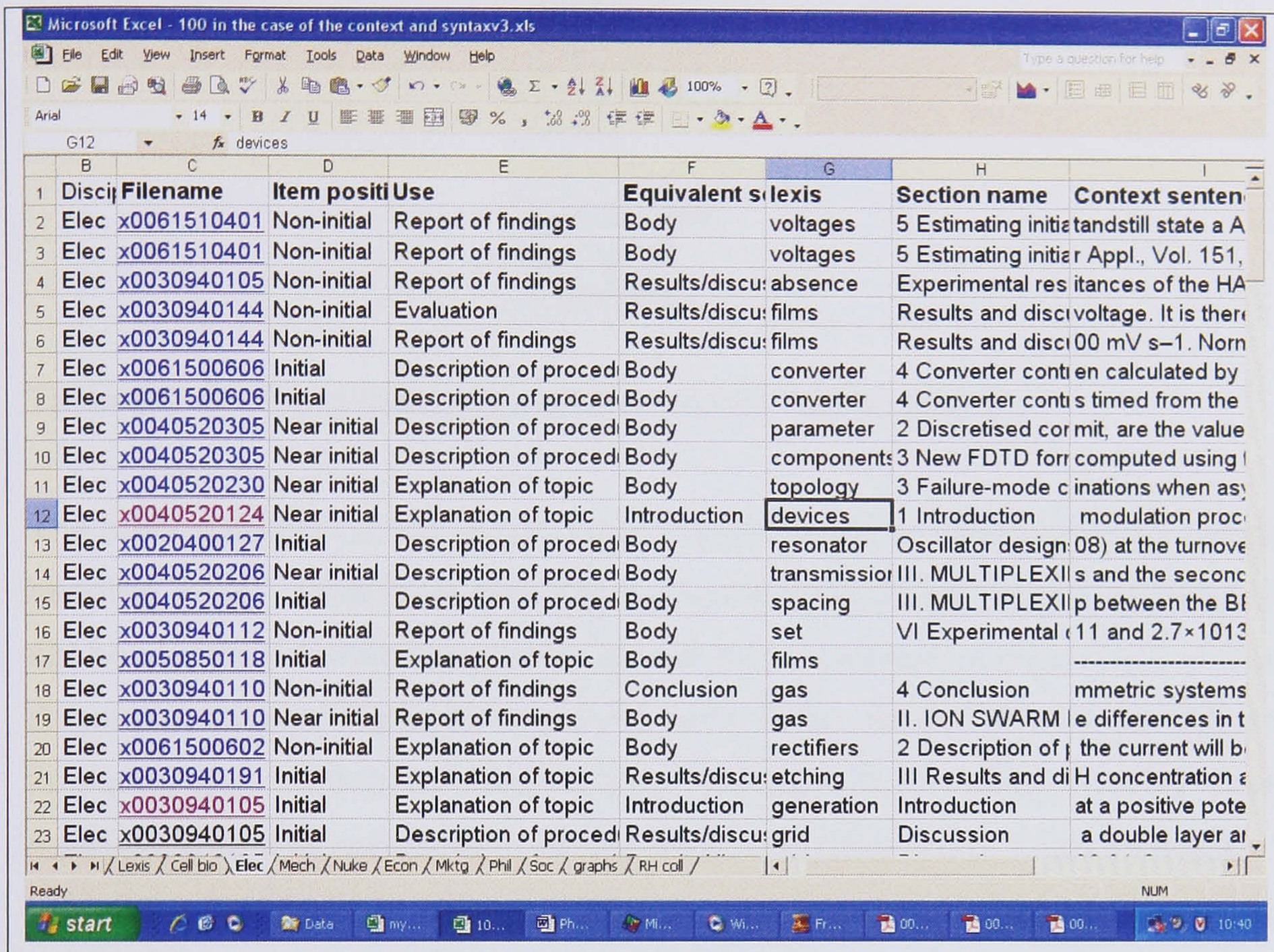


Figure 5.13: Identification of the textual location of *in the case of the* using the search function of Adobe Acrobat Reader 8

Figure 5.13 above shows the same pattern identified in the same article displayed in pdf format. A comparison of Figure 5.12 and Figure 5.13 shows the advantages of pdf over plain text format. The sub-heading "Introduction" is visible in the same screen as the pattern in the pdf version, making obvious its location in the article. The researcher needs to scroll further through the plain text version of this article in order to find the pattern; extra-textual noise from a footer interrupts the text in the top right-hand portion of the figure.

An Excel file was used to facilitate the qualitative investigation of patterns. The file, as shown in Figure 5.14 contained eight worksheets, one for each discipline. Each worksheet contained one row for each occurrence of the pattern, and each row consisted of cells containing information from the analysis of the pattern in its wider context. Two cells in each row indicated the pattern's position in the sentence and the section of the article in which it occurred. The adjacent cell contained this section's equivalent in the discourse macro-structure, which allowed this discipline to be compared with other disciplines. For example, the sections of Electrical Engineering articles with the different titles *Experimental Results*, *Results and Discussion*, *Discussion*, and *Numerical Results* were all classed as being Results /Discussion. This would allow pattern locations to be compared with those from other disciplines where sections containing results might have slightly different titles. Articles in Sociology, for example, present results in sections with titles such as *Findings, Discussion, Results, Empirical Evidence, or Analysis*. Other cells in each row contained the use identified for this pattern, the headword of the noun group immediately to the right of the pattern, and the context sentence from the concordance line. Finally a hyperlink was created from the cell containing the filename of the article to the pdf file so the original context could be accessed quickly, as in Figure 5.13 above.



	B	C	D	E	F	G	H	I
1	Disci	Filename	Item positi	Use	Equivalent s	lexis	Section name	Context senten
2	Elec	x0061510401	Non-initial	Report of findings	Body	voltages	5 Estimating initia	tandstill state a A
3	Elec	x0061510401	Non-initial	Report of findings	Body	voltages	5 Estimating initia	r Appl., Vol. 151,
4	Elec	x0030940105	Non-initial	Report of findings	Results/discus	absence	Experimental res	itances of the HA
5	Elec	x0030940144	Non-initial	Evaluation	Results/discus	films	Results and disc	voltage. It is there
6	Elec	x0030940144	Non-initial	Report of findings	Results/discus	films	Results and disc	00 mV s−1. Norn
7	Elec	x0061500606	Initial	Description of proced	Body	converter	4 Converter conti	en calculated by
8	Elec	x0061500606	Initial	Description of proced	Body	converter	4 Converter conti	s timed from the
9	Elec	x0040520305	Near initial	Description of proced	Body	parameter	2 Discretised cor	mit, are the value
10	Elec	x0040520305	Near initial	Description of proced	Body	components	3 New FDTD forr	computed using
11	Elec	x0040520230	Near initial	Explanation of topic	Body	topology	3 Failure-mode c	inations when as
12	Elec	x0040520124	Near initial	Explanation of topic	Introduction	devices	1 Introduction	modulation proc
13	Elec	x0020400127	Initial	Description of proced	Body	resonator	Oscillator design	08) at the turnove
14	Elec	x0040520206	Near initial	Description of proced	Body	transmissior	III. MULTIPLEXI	s and the seconc
15	Elec	x0040520206	Initial	Description of proced	Body	spacing	III. MULTIPLEXI	p between the BI
16	Elec	x0030940112	Non-initial	Report of findings	Body	set	VI Experimental	11 and 2.7×1013
17	Elec	x0050850118	Initial	Explanation of topic	Body	films		----------
18	Elec	x0030940110	Non-initial	Report of findings	Conclusion	gas	4 Conclusion	mmetric systems
19	Elec	x0030940110	Near initial	Report of findings	Body	gas	II. ION SWARM	e differences in t
20	Elec	x0061500602	Non-initial	Explanation of topic	Body	rectifiers	2 Description of	the current will b
21	Elec	x0030940191	Initial	Explanation of topic	Results/discus	etching	III Results and di	H concentration a
22	Elec	x0030940105	Initial	Explanation of topic	Introduction	generation	Introduction	at a positive pote
23	Elec	x0030940105	Initial	Description of proced	Results/discus	grid	Discussion	a double layer ar

Figure 5.14: Excel file used to facilitate the qualitative investigation of *in the case of the*

Determining the location in text for journal articles in different disciplines is problematic. The names given to sections in research articles vary between disciplines, between journals in the same discipline, and even between articles in the same journal. Some journals distinguish between a “results” section and a “discussion” section, while others have a single section called “results and discussion”. Some sections in Philosophy articles are numbered without headings, whereas Cell Biology article sections generally have the same sequence of headings.

This problem can be addressed by developing a standard macro-structure of headings, which can be developed from a sample of the different journals, as shown in Table 5.5 below:

Abstract
Introduction
Methodology
Body
Results/discussion
Conclusion
Appendix
Footnote

Table 5.5: Standard macro-structure of journal article sections

The main drawback to the above macro-structure is that it features a section with the heading "Body", containing all those text sections for which a common heading could not be identified. While this model represents only a small improvement on the discourse model of Nattinger and DeCarrico (1992), which Willis (1995: 89) criticised for seeming to "lead to the disappointing conclusion that essays, informal letters, and business letters all consist of opening, body, and conclusion," it nonetheless provides a framework in which comparison can be made between stages of the genre of the research article in different disciplines.

5.6 Summary

This chapter has discussed the reasons for the methodological procedures chosen for this study. It outlined approaches taken in previous corpus-based and corpus-driven studies of fixed collocational patterns in corpora of written academic English and showed why it was necessary to construct a corpus for the present study. It has discussed in detail the composition of the corpus, the procedures used to ensure that its subcorpora were comparable, and the measures taken to allow for noisy data. The chapter has also described the methodological procedures for identifying the forms and textual locations of the fixed collocational patterns in the different versions of the corpus. We now move on to the quantitative results of this study.

Chapter 6 Quantitative results

6.1 Introduction

This chapter presents the results of searches for fixed collocational patterns in the research article corpus. For both isolexical and isotextual version of the corpus it describes the overall distribution of these forms in the corpus and then in individual disciplines. It then presents the fixed collocational patterns which are found to occur in all eight disciplines.

6.2 Isolexical and isotextual frequency comparison: a brief example

The first quantitative finding from this study is that there are very few 6, 7, or 8-word fixed collocational patterns which occur more than 100 times and which are part of running text. All 7 and 8-word clusters which occur 100 times or more consist of extra-textual noise, such as *may print download or email articles for individual* or *refine search print e-mail save folder is empty*. Of those 7 and 8-word clusters from running text, none occurs more than 100 times; indeed nearly all occur fewer than ten times and in only one text. Examples of these in Philosophy are *i am not a brain in a vat*, which occurs 40 times in 5 texts, *i am thinking that some water is dripping* (40 times in 2 texts) and *the sound exercise of a capacity for practical* (21 times in one text).

The frequencies for 6-word patterns are similarly low, with the exception of two clusters, *in the presence or absence of* and *as described in materials and methods* which meet the frequency criteria for fixed collocational patterns in both the isolexical and isotextual versions of the corpus. Since such patterns do not occur in every discipline, however, they cannot be discussed as part of the main quantitative results of this study. Despite this, these occurrences are useful as a way of briefly illustrating the methodology adopted to compare the occurrence of fixed collocational patterns between isolexical and isotextual versions.

Table 6.1 below shows the 6-word clusters which meet the isolexical fixed collocational pattern frequency threshold of at least 100 occurrences. It shows the isolexical and isotextual frequency of each cluster, and also their rank in each frequency distribution. It can be seen that only the first two clusters in Cell Biology occur isotextually more than 100 times, and can thus be considered as fixed collocational patterns. The right-hand column is a measure of the difference in ranking in the frequency distribution of a cluster between the two versions. If a cluster's isolexical rank is the same as its isotextual rank, then the rank difference is equal to zero. It can be seen from Table 6.1 that, despite the difference between their isolexical and isotextual frequencies of occurrence, all 6-word clusters have a rank difference of zero since the most frequent 6-word cluster in the isolexical subcorpus in each discipline is also the most frequent 6-word cluster in the isotextual subcorpus. The second most frequent 6-word cluster in the isolexical subcorpus in each discipline is also the second most frequent 6-word cluster in the isotextual subcorpus, and so on.

Discipline/6-word pattern or cluster	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
Cell Biology					
in the presence or absence of	265	1	198	1	0
as described in materials and methods	246	2	197	2	0
in the absence or presence of	109	3	84	3	0
Electrical Engineering					
it can be seen that the	171	1	65	1	0
it should be noted that the	118	2	44	2	0
Mechanical Engineering					
it can be seen that the	154	1	86	1	0
it should be noted that the	110	2	66	2	0
Philosophy					
from the point of view of	106	1	86	1	0

Table 6.1: Comparison of isolexical and isotextual frequencies of 6-word fixed collocational patterns

The rank difference is a useful measure since it indicates the frequency of a pattern relative to the frequencies of other patterns, regardless of what its actual frequency is in either the isolexical and isotextual version. If a pattern has a low (i.e. near 0)

rank difference, then its rank is very similar in both versions of the subcorpus and reflects the stability of the distributions of patterns in the two versions. A positive rank difference for a pattern would mean that it has a higher frequency rank in the isotextual version than in the isolexical version, while a negative frequency rank difference would mean a higher isolexical frequency rank.

It is also revealing to compare the rank difference with the VSD measure described earlier. It would be expected that most patterns would have a low rank difference in a subcorpus with a higher VSD. In Marketing, for example, which has a VSD of 1, the ranking of patterns would be expected to be similar in both the isolexical and isotextual subcorpora, since the texts and tokens in both versions are virtually identical. It would also be expected that patterns would have a higher rank difference in a subcorpus with a lower VSD, such as Electrical Engineering which has a VSD of 0.34. This would be because there are far fewer texts and tokens in the isotextual version than in the isolexical version, and so it could be expected that patterns which are isolexically frequent may not be so frequent isotextually. This in turn would mean that the change in frequency would be so large as to reduce their ranking in the isotextual frequency distribution, and so lead to a higher rank difference for that pattern between the versions.

It can be seen in the brief example in Table 6.1, however, that the 6-word clusters (still called *clusters* here because isotextually they occur too infrequently to be considered fixed collocational patterns) in Marketing have a rank difference of zero, which is not surprising, but 6-word clusters in Electrical Engineering also have a rank difference of zero, which is contrary to what might be expected since its VSD is much lower. This type of table is used later in this section to show the rank differences between 3-, 4- and 5-word fixed collocational patterns in each discipline.

Having presented the rather sparse results for 6-, 7- and 8-word clusters, but given in the process an idea of the usefulness of the two measures VSD and rank difference, this chapter continues with the results for 3-, 4-, and 5-word patterns in general.

6.3 Frequency distributions for fixed collocational patterns

Table 6.2 below shows the distribution of occurrences of 3-, 4-, and 5-word fixed collocational patterns in the isolexical and isotextual versions of the research corpus. These results give an indication of the differing levels of patterning in different disciplines, and between the isolexical and isotextual versions of the corpus. Figures are given for types, i.e. different patterns, and tokens, i.e. the number of times each pattern occurs. While all these types and tokens were originally found as clusters, they can all be considered to be fixed collocational patterns: the table shows only those forms remaining after the removal of noisy clusters (see Table 5.4 above) and the implementation of the cut-off point of 100 occurrences (equivalent to 20 times per million tokens in the isolexical version of the corpus, and 20 times per hundred texts in the isotextual version). Table 6.2 shows the number of patterns found in each version of the corpus as a whole, and also in each academic discipline.

		3-word patterns		4-word patterns		5-word patterns	
Version	Subcorpus (VSD)	Isolexical	Isotextual	Isolexical	Isotextual	Isolexical	Isotextual
	Full corpus	5,931	2,446	800	325	78	30
	Cell Bio. (0.8)	1,192	855	157	96	13	5
	Elec. Eng. (0.34)	1,469	285	204	30	35	3
	Mech. Eng. (0.52)	1,364	489	166	62	19	4
Types	Nucl. Phys. (0.54)	1,700	640	273	81	20	3
	Economics (1.09)	1,135	1,310	122	153	6	9
	Marketing (1.00)	1,017	1,008	90	88	4	4
	Philosophy (0.94)	1,452	1,346	192	174	13	9
	Sociology (1.05)	867	896	74	79	3	3
	Full corpus	2,179,950	1,405,945	244,508	138,074	18,345	5,970
	Cell Bio. (0.8)	244,826	169,022	28,595	17,177	2,129	734
	Elec. Eng. (0.34)	324,004	52,689	45,197	6,297	6,306	464
	Mech. Eng. (0.52)	290,408	93,889	34,236	11,085	3,026	585
Tokens	Nucl. Phys. (0.54)	370,737	126,299	50,924	14,345	3,485	708
	Economics (1.09)	239,996	277,160	22,531	28,187	683	1,182
	Marketing (1.00)	206,952	204,800	16,808	16,779	704	743
	Philosophy (0.94)	323,773	294,545	33,652	29,963	1,655	1,197
	Sociology (1.05)	179,942	187,894	14,677	14,679	357	357

Table 6.2: Frequencies of 3, 4, and 5-word pattern types and tokens occurring 100 times or more in the isolexical and isotextual versions of the corpus

It can be seen that in all disciplines, in both versions of the corpus, there are many more 3-word combination tokens than 4 and 5-word combination tokens, and many more 4-word combination tokens than 5-word combination tokens. This is a common property of fixed collocational patterns in general, as discussed in work on lexical bundles (Biber et al. 1999; Cortes 2004; Hyland 2008) and is inevitable, given that 5 word patterns are made up of 4 and 3-word patterns, and 4-word patterns are made up of 3-word patterns and so on. The VSD, the difference in size (in word tokens) between the isolexical subcorpus of a discipline in relation to its isotextual subcorpus, is also shown for each discipline.

6.3.1 Isolexical and isotextual frequency distributions of fixed collocational pattern types in the research article corpus

An isolexical comparison between the disciplines reveals that the discipline with the largest number of 3-word patterns is Nuclear Physics, followed by Electrical Engineering, Mechanical Engineering, and Cell Biology. It would thus be reasonable to conclude from this observation that, since 3-word patterns are more frequent in 5 million words of these four disciplines than in 5 million words from the other disciplines, Science research articles contain more patterning than those in non-Science disciplines. However, this conclusion would be based on the assumption that 5 million words in each discipline contain not only the same amount of language but also the same amount of communication. In fact, as discussed in the previous chapter, this conclusion is based on a biased corpus which favours the disciplines containing shorter texts. If the isotextual frequencies of these patterns are compared, then this bias is eliminated.

Table 6.2 therefore reveals much about how the elimination of this bias affects the frequencies of different patterns. While there are more 3, 4, and 5-word pattern types in Science disciplines than in non-Science disciplines in the isolexical version of the corpus, this distribution is reversed in the isotextual version of the corpus, so that there are more 3, 4, and 5-word pattern types in non-Science disciplines than in Science disciplines. This distinction is more noticeable for 3-word combinations than for 4 and 5-word combinations, as shown graphically in Figure 6.1 to Figure 6.3 below.

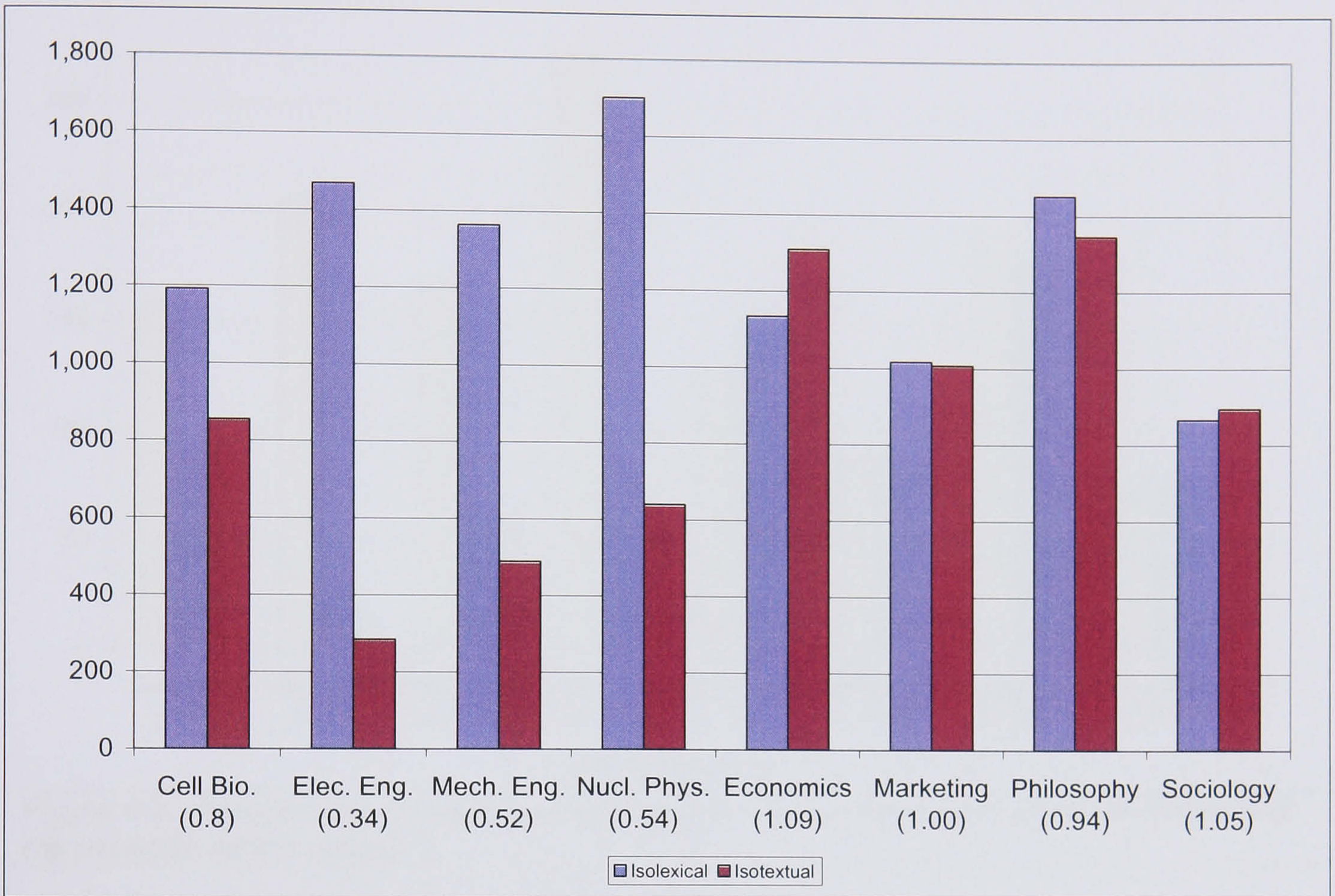


Figure 6.1: Frequencies of 3-word pattern type in the isolexical and isotextual versions of the research article corpus

When the same amount of language is compared, fixed collocational patterns are more common in Science disciplines, but the reverse is true when the same amount of communication is compared. This suggests that the motivation for fixed collocational patterns to occur frequently isolexically depends on the number of texts in the corpus. It also suggests that this motivation may not have been accounted for in previous isolexical corpus-based studies of collocation.

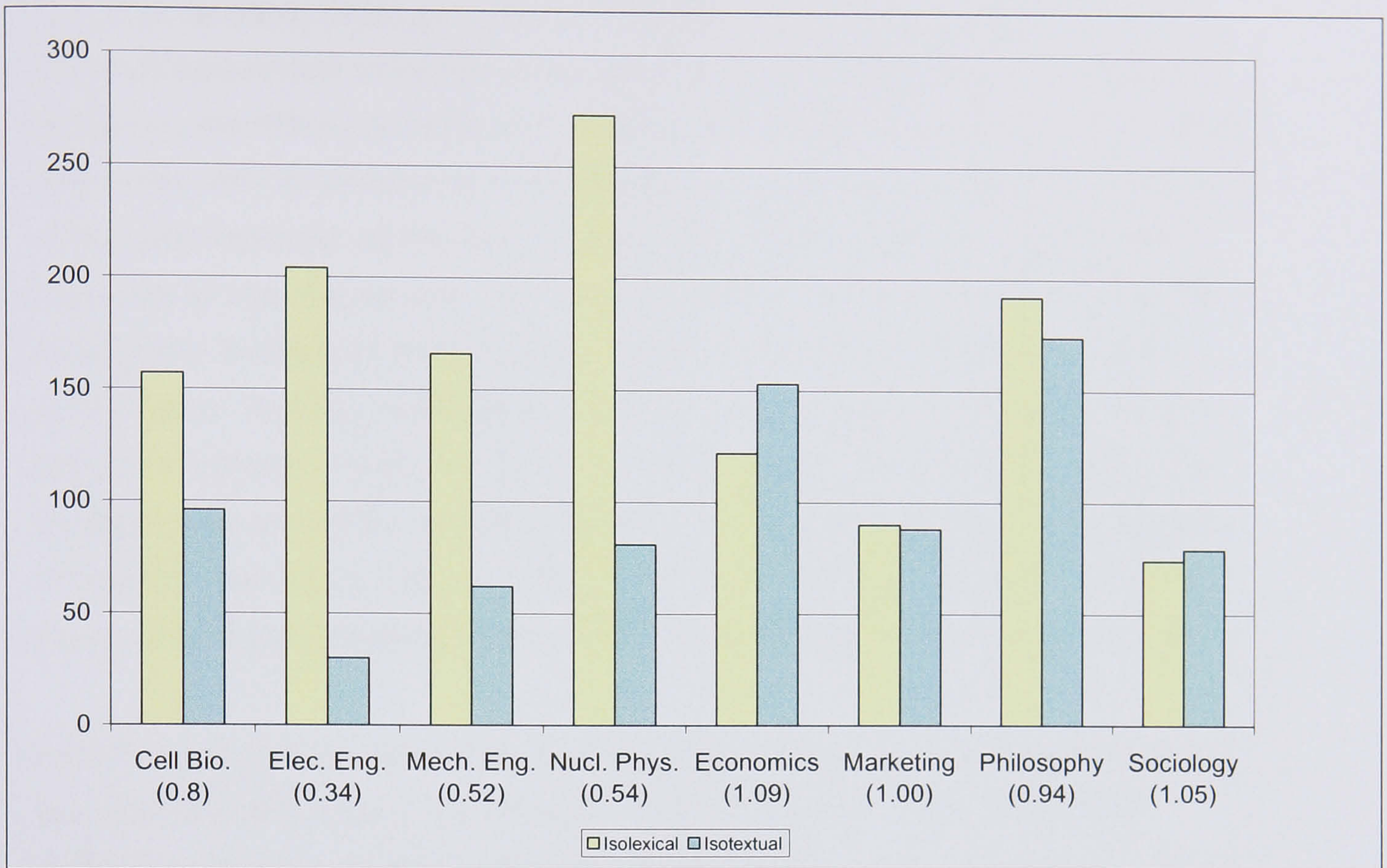


Figure 6.2: Frequencies of 4-word pattern types in the isolexical and isotextual versions of the research article corpus

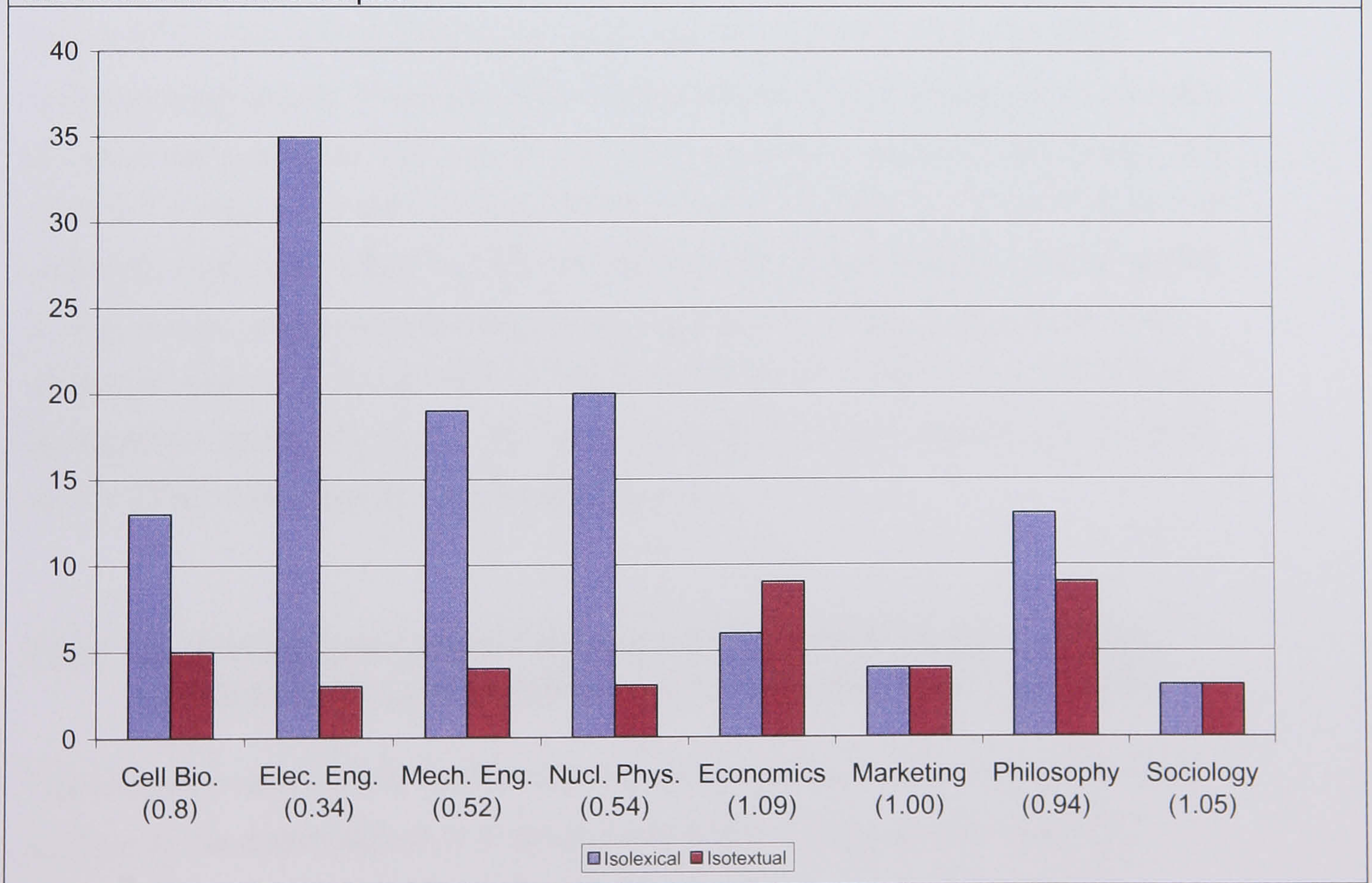


Figure 6.3: Frequencies of 5-word pattern types in the isolexical and isotextual versions of the research article corpus

One other tendency observed concerns the differences between the isolexical and isotextual frequencies of pattern types as can be seen in Table 6.2. In non-Science disciplines, the difference between isolexical and isotextual frequencies by and large reflects the VSD for that discipline. In Marketing, for example, in which the isolexical version and isotextual version are the same size, giving a VSD of 1, pattern types occur with similar frequencies: 1,017 and 1,008 in the case of 3-word pattern types, 90 and 88 in the case of 4-word pattern types, and 4 and 4 in the case of 5-word pattern types. Similarly in Philosophy, in which the isotextual version is 94% of the size of the isolexical version (giving a high VSD of 0.94), the isotextual frequencies of patterns are smaller by a similar proportion: 1,452 and 1,346 (92.6%) in the case of 3-word pattern types, 192 and 174 (90.6%) in the case of 4-word pattern types, and a rather larger difference 13 and 9 (69.2%) in the case of 5-word pattern types.

By contrast, in Science disciplines the difference between the two frequencies does not reflect the difference in size between the two versions, i.e. the difference is greater than the VSD might suggest. In Electrical Engineering, for example, in which the isotextual version is roughly a third of the size of the isolexical version (a low VSD of 0.34), the isotextual frequencies are less than a third of the size of the isolexical frequencies: 1,469 and 285 (19.4% of the isolexical frequency) in the case of 3-word pattern types, 204 and 30 (14.7%) in the case of 4-word pattern types, and 35 and 3 (8.5%) in the case of 5-word pattern types. Similarly for Nuclear Physics, in which the isotextual version is just over half the size of the isolexical version (giving a VSD of 0.54), the isotextual frequencies are less than half as frequent as in the isolexical version: 1,700 and 640 (37.6% of the isolexical frequency) in the case of 3-word pattern types, 273 and 81 (29.7%) in the case of 4-word pattern types, and 20 and 3 (15%) in the case of 5-word pattern types.

6.3.2 Isolexical and isotextual frequency distributions of fixed collocational pattern tokens in the research article corpus

The tendency observed for fixed collocation pattern *types* in section 6.3.1 can also be seen in the distribution of 3, 4, and 5-word pattern *tokens* in each discipline. Patterns are again more common isolexically in Science disciplines, but more common isotextually in non-Science disciplines, as shown in Figure 6.4 to Figure 6.6

below. Again this tendency would have been missed in a purely isolexical comparison.

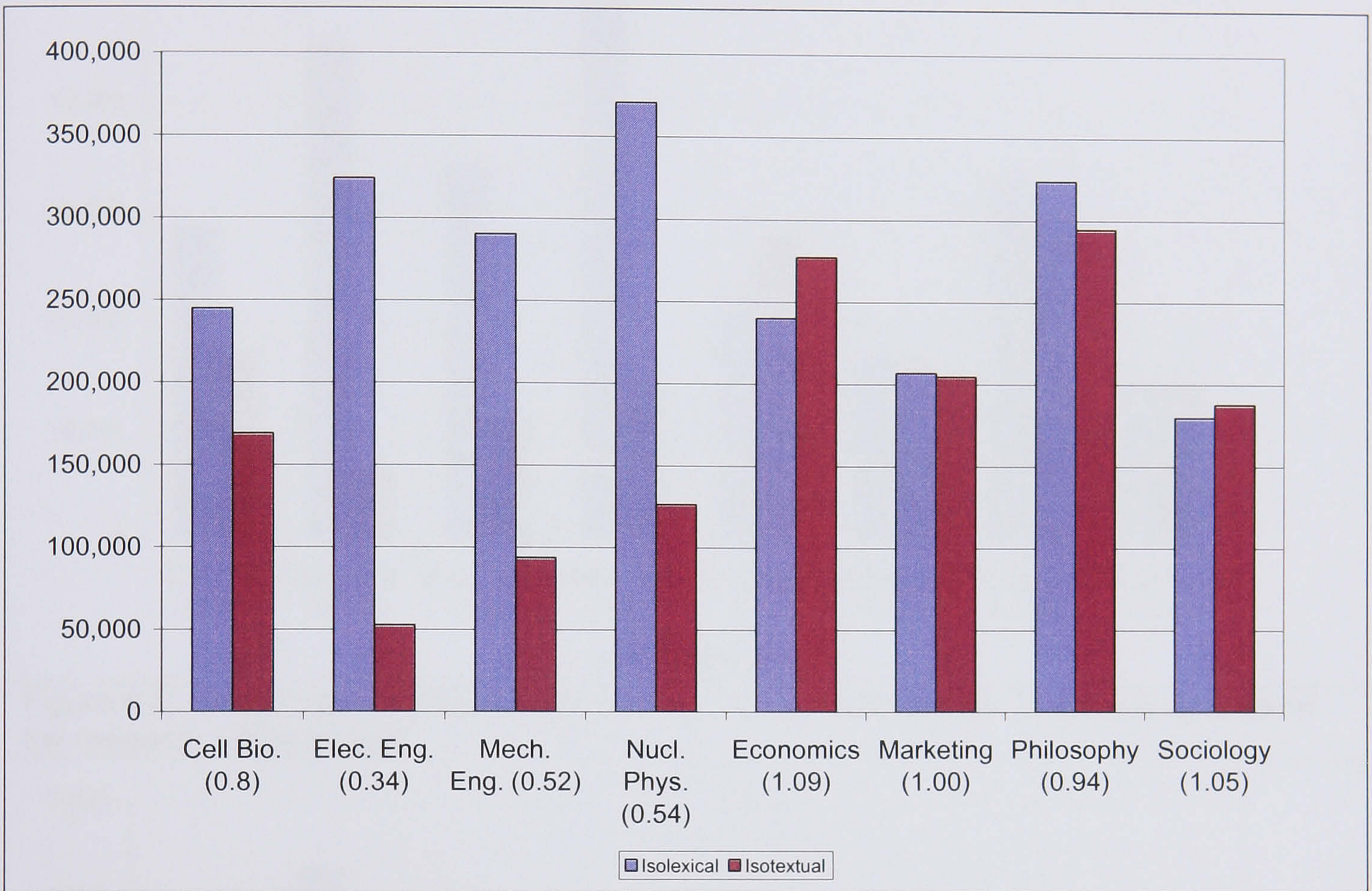


Figure 6.4: Frequencies of 3-word pattern tokens in the isolexical and isotextual versions of the research article corpus

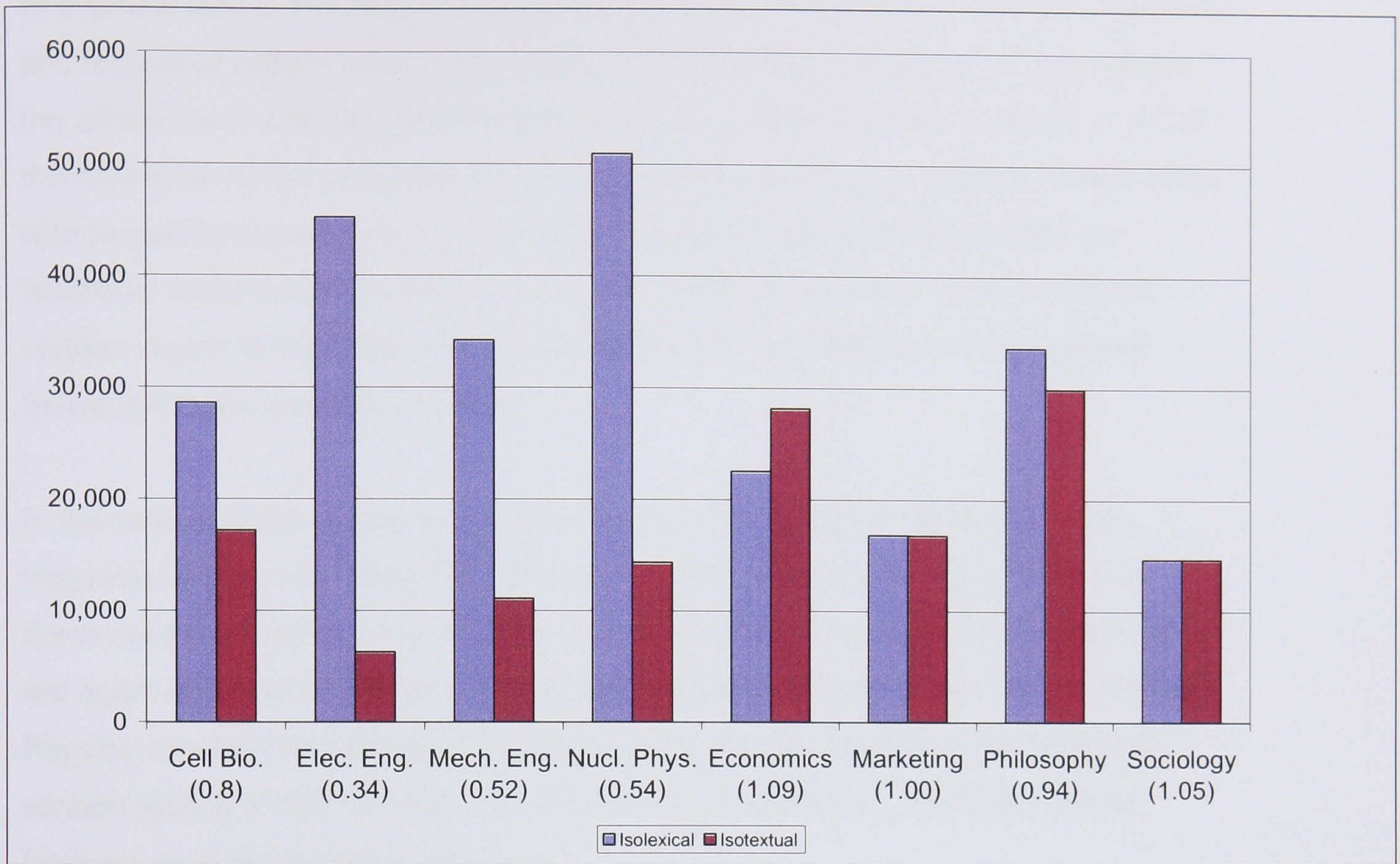


Figure 6.5: Frequencies of 4-word pattern tokens in the isolexical and isotextual versions of the research article corpus

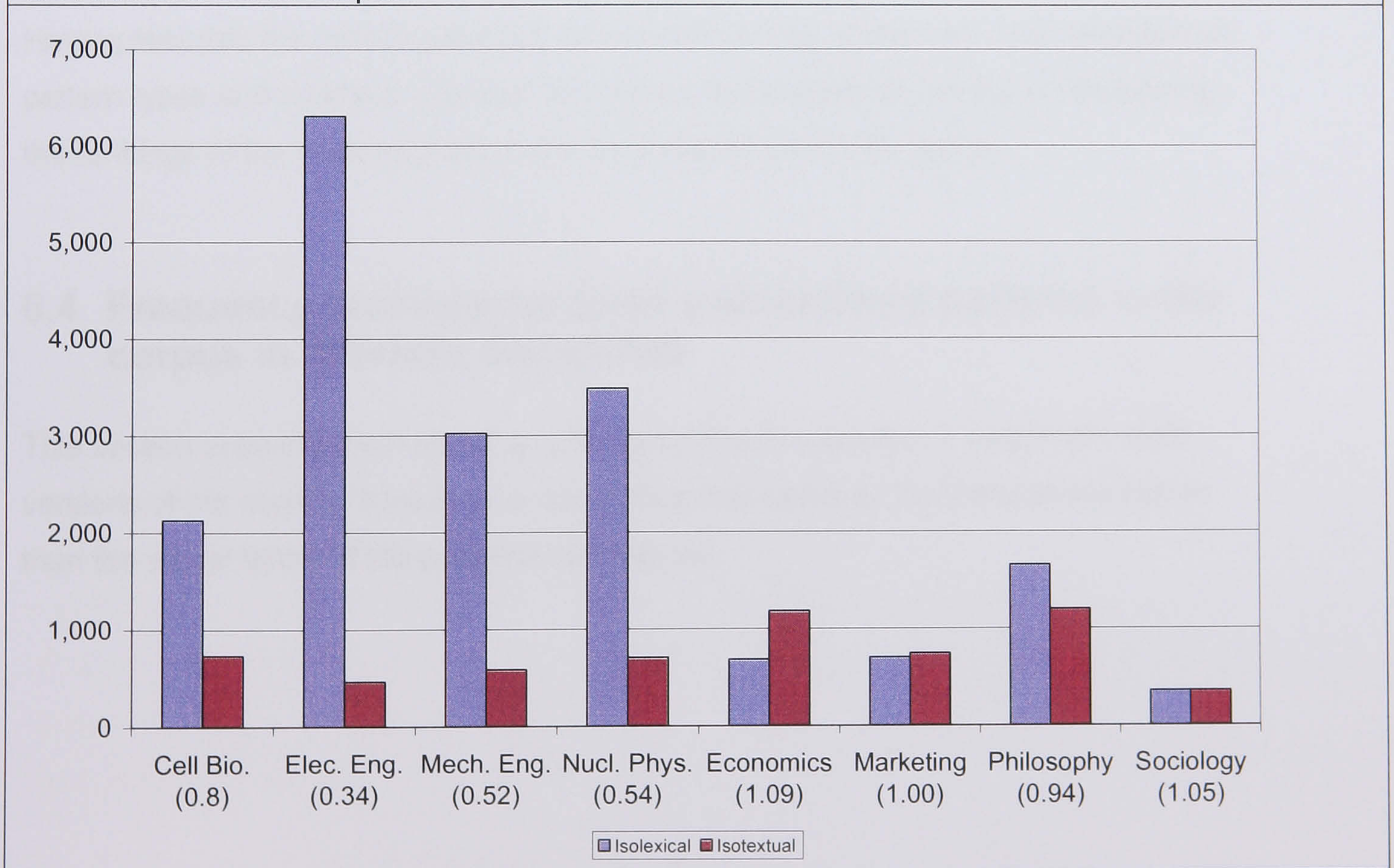


Figure 6.6: Frequencies of 5-word pattern tokens in the isolexical and isotextual versions of the research article corpus

In a similar way to the frequencies of pattern types, the difference between isolexical and isotextual pattern token frequencies in non-Science disciplines largely reflects the difference in size between the two versions. In Marketing, for example, in which the isolexical version and isotextual version are the same size, pattern tokens occur with similar frequencies, as can be seen in Table 6.2, while in Philosophy the isotextual frequencies of pattern tokens reflect the smaller size of the isotextual version. Again in the case of 5-word pattern types the difference in frequencies between the two versions is larger.

In the case of Science disciplines, once more, the difference between the two frequencies does not reflect the difference in size between the two versions. In Electrical Engineering, for example, with a VSD of 0.34, the isotextual frequencies are again less than a third as frequent as the isolexical frequencies, and in Nuclear Physics, in which the isotextual version is just over half the size of the isolexical version (giving a VSD of 0.53), the isotextual frequencies are less than half as frequent as in the isolexical version.

Having reported the overall isolexical and isotextual frequencies of fixed collocational pattern types and tokens in different disciplines, the chapter continues by discussing the rankings of the most frequent patterns to occur in each discipline.

6.4 Frequency rankings for fixed collocational patterns in the corpus in different disciplines

This section presents the frequency rankings of fixed collocational patterns in both versions of the corpus. It focuses on the differences between the frequencies rather than the actual forms of the patterns themselves.

6.4.1 Isolexical and isotextual frequency rankings for the top 50 3-word fixed collocational patterns in different disciplines

This section presents the frequencies of 3-word patterns in the corpus in each discipline. Figure 6.7 and Figure 6.8 below show a graphical representation of the isotextual and isolexical frequencies of 3-word patterns in the corpus in each discipline. This gives an idea of the distribution of patterns in each discipline, such as how much more frequently the most frequent pattern occurs than the next most frequent, and how much more frequently in turn this pattern is to the next most frequent and so on. The figures also allow the distributions for the disciplines to be seen in relation to each other. The coloured lines are shown to aid the eye only: the line between each value does not represent a frequency of occurrence between two patterns. Figure 6.7 shows the differences between disciplines in an isolexical comparison, while Figure 6.8 shows the differences in an isotextual comparison. The distributions are also shown in tabular form in Appendices 4-11, which also show which pattern occurs in each rank in the distribution.

The isolexical comparison between the disciplines shows that the most frequent 3-word pattern occurs in Electrical Engineering, as shown in Figure 6.7 below. This pattern is *shown in fig* (see Appendix 5) and it occurs isolexically far more than the next most frequent pattern in that discipline or any other. The next most frequent 3-word pattern is in Mechanical Engineering, and is again *shown in fig* (see Appendix 6), and the next most frequent is *the presence of* in Cell Biology (see Appendix 4). The pattern frequency curve of these disciplines is initially very steep, as the most frequent pattern occurs many times more than the next pattern, which in turn occurs many times more frequently than the next pattern, and so on. After the first five patterns there is less and less difference between the frequencies of successive patterns in the list, and so the frequency curves flatten rapidly. This gives the isolexical frequency distributions for 3-word patterns in these three disciplines their distinctive “long tail” shape, so that after the first 30 patterns or so their frequencies have largely converged.

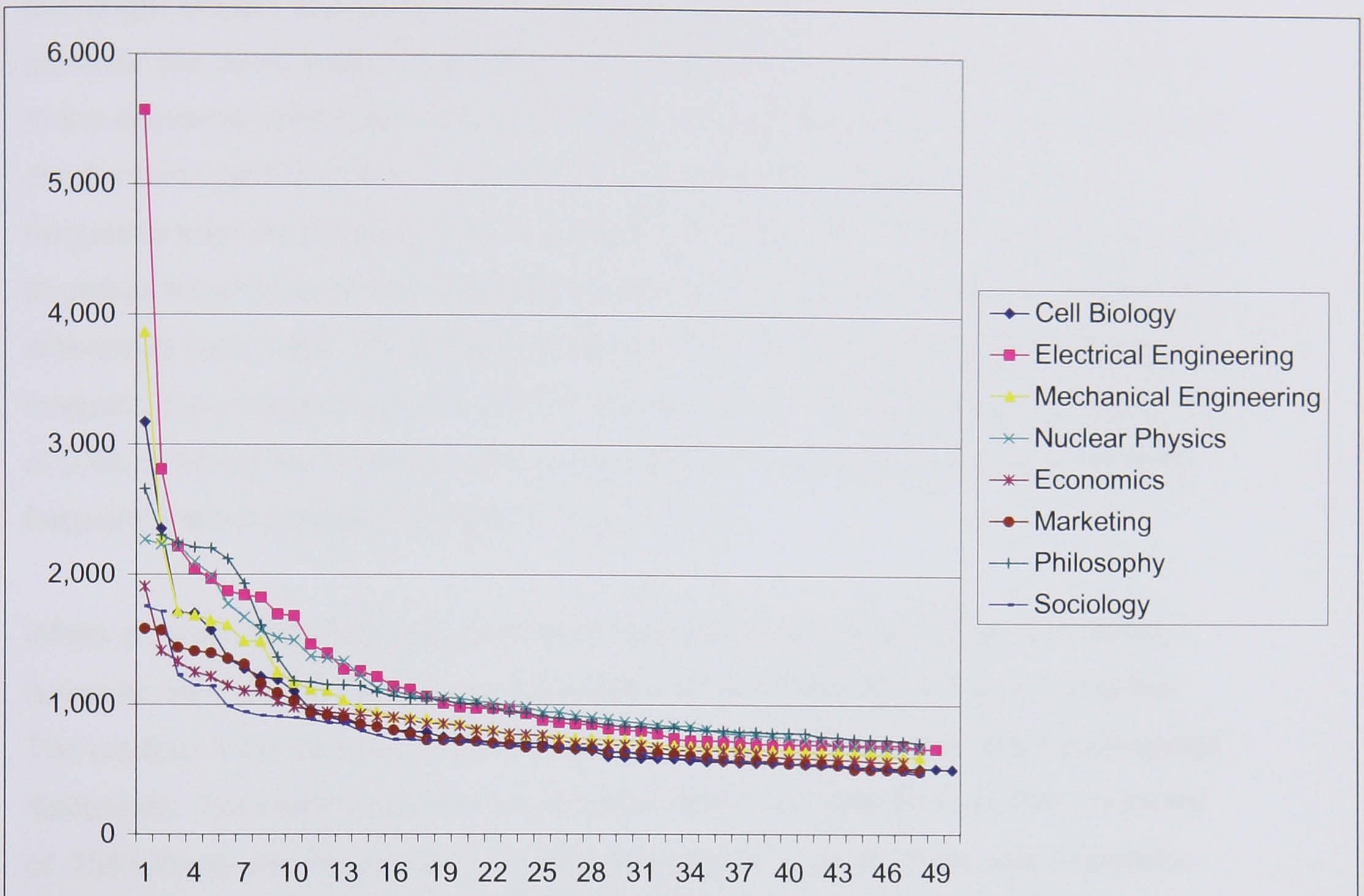


Figure 6.7: Isolexical frequency distribution of the top 50 3-word patterns

After Electrical Engineering, Mechanical Engineering, and Cell Biology, the discipline in which the most frequent 3-word pattern is the next most frequent is Philosophy where *that it is* occurs 2660 times (see Appendix 10). This is followed by Nuclear Physics (*in order to,* 2270 times; see Appendix 7), Economics, (*the number of,* 1910 times; see Appendix 8), Sociology (*in terms of,* 1757 times; see Appendix 11), and finally Marketing (*the number of*, 1585 times; see Appendix 9). The frequency distributions for these latter five disciplines are flatter, being less steep initially, and gradually begin to follow the distributions for the first three disciplines. The distributions are mostly concave, although the distribution for Philosophy patterns is convex from the third to the tenth point. This feature occurs when adjacent patterns in a distribution have similar frequencies. After 30 patterns the frequency curves for all disciplines are almost identical in shape, with the difference in frequencies between them remaining largely constant.

These isolexical frequency distributions suggest that there is a relationship between the length of texts in a discipline and the frequency of the most frequent 3-word patterns: the more texts, the greater the number of occurrences of 3-word patterns. In the isolexical comparison the most frequent 3-word patterns in the disciplines with shorter texts (and therefore more texts in their subcorpus) occur much more frequently than the frequent 3-word patterns in the disciplines with longer texts (and therefore fewer texts in their subcorpus). It appears that the greater the version size differential (see Table 5.3 above), the greater the frequency with which the most frequent 3-word pattern occurs. The one exception to this is Philosophy, with a VSD of 0.94, in which the most frequent 3-word pattern is more frequent than the most frequent 3-word pattern in Nuclear Physics (0.54).

When a comparison between isotextual frequencies of 3-word patterns is made, however, several differences are noticeable, as can be seen in Figure 6.8 below. The top four most frequent 3-word patterns occur in both Science and non-Science disciplines. The most frequent 3-word patterns occur in Cell Biology (*the presence of*, 2564 times; see Appendix 4), Philosophy (*that it is,* 2468 times; see Appendix 10), Economics (*the number of*, 2147 times; see Appendix 8) and Mechanical Engineering (*shown in fig*, 2001 times; see Appendix 6). The frequency curves again follow a long tail distribution, although Philosophy again has a convex shape initially. After 40 patterns the frequency curves for all disciplines are again almost identical in shape, with the difference in frequencies between them remaining largely constant.

The main finding of this first comparison of individual disciplines is that the frequency of occurrence of particular 3-word patterns in different disciplines therefore depends on the version of the corpus which is used to make the comparison. In the isolexical comparison, 3-word patterns occur more in disciplines containing a larger number of shorter texts than in disciplines containing a smaller number of longer texts. This finding suggests that 3-word patterns may have a functional role, as there appears to be a relationship between the number of texts and the frequency of these patterns.

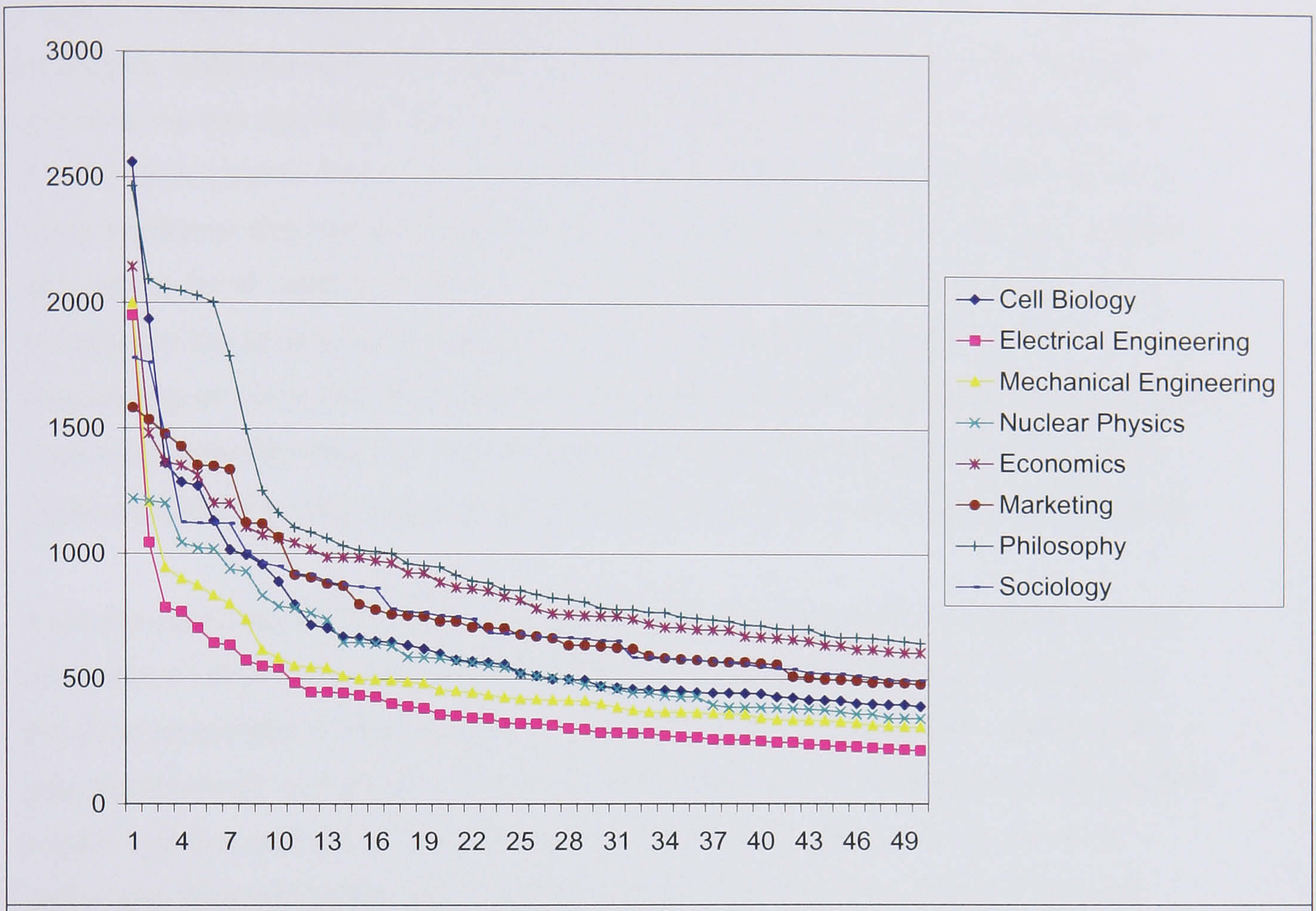


Figure 6.8: Isotextual frequency distribution of the top 50 3-word patterns

6.4.2 Rank differences for the top 50 3-word fixed collocational patterns in different disciplines

The next comparison needs to be made between the actual forms of the 3-word patterns in the frequency tables for each discipline. While the previous section has revealed that there is clearly a difference between isolexical and isotextual frequencies of 3-word patterns in each discipline in the corpus, it is not clear whether these patterns are the same in each version of the corpus. The most frequent 3-word pattern isolexically in one discipline may not be the most frequent pattern isotextually.

Table 6.3 below summarises the rank differences between isolexical and isotextual frequency distributions for the top 50 3-word patterns in each discipline. Its eight columns are the right-hand column taken from each of the tables in Appendices 4-11. The third column from the left, for example, contains the rank differences for 3-word bundles in Mechanical Engineering calculated in Appendix 6. Putting the rank differences for all eight disciplines in one table allows the rank differences in each discipline to be compared at a glance. It can be seen that the difference in frequencies of 3-word bundles between the isolexical and isotextual versions of each subcorpus has little effect on the frequency rankings of these patterns: the 3-word patterns which are most frequent isolexically are also the most frequent isotextually.

It can be seen that in the top 20 rank positions in all subcorpora no patterns have a rank difference of more than 5 places. This is regardless of the VSD: for example, the most isolexically frequent pattern in Electrical Engineering (0.34), *shown in fig* (see Appendix 5), is also the most isotextually frequent, i.e. the pattern is ranked first in both versions and so its rank difference is zero. In Marketing (1.00) the most isolexically frequent pattern, *the number of* (see Appendix 9), is likewise also the most isotextually frequent. Even though the two versions of the Marketing subcorpus are the same size, however, they do not contain identical texts, and so the rankings are not completely stable.

Lower down the frequency table, rank differences become larger as the tail of the distribution is reached. At this point in the distribution patterns occur with very similar frequencies, so a small change in the frequency of a pattern can have a larger effect on its ranking in the frequency table, making the rankings of the patterns between the two versions less stable overall.

It is noticeable that one pattern in Electrical Engineering has a rank difference which stands out from the others. This is *in this letter* which, as shown in Appendix 5, occurs far less often isotextually (ranked 90th with 150 occurrences) than it does isolexically (ranked 23rd with 971 occurrences), and so has a rank difference of -67. This does not appear to be an idiosyncratic use: it occurs in 790 texts isolexically and in 153 texts in the isotextual subcorpus, and so is not used repeatedly by a small number of writers of a small number of texts.

Cell Biology (0.80)	Electrical Engineering (0.34)	Mechanical Engineering (0.52)	Nuclear Physics (0.54)	Economics (1.09)	Marketing (1.00)	Philosophy (0.94)	Social Science (1.05)
0	0	0	0	0	0	0	0
0	0	0	-4	0	0	0	0
-1	-1	-3	0	-1	0	0	0
1	1	1	-3	1	0	-2	-3
0	-4	1	3	0	-2	1	-1
0	-4	1	1	-1	1	1	-2
0	2	0	-1	-1	1	0	-3
0	2	0	4	2	-1	0	4
0	2	0	-3	-1	1	0	4
0	2	0	1	-1	0	-2	-4
-1	0	-1	-2	-1	0	0	-4
1	0	1	1	3	0	2	-1
0	1	0	3	-2	0	-4	2
-2	-2	0	-5	1	0	1	5
0	1	-1	1	1	-2	1	-4
2	1	-6	2	0	1	0	0
0	4	-1	1	0	1	-2	0
-1	1	-2	3	-1	-5	3	-5
-3	-1	2	-9	-3	-1	-1	1
0	-8	-4	-3	2	1	2	-1
0	3	2	0	-2	3	0	-8
4	-3	-4	5	2	1	0	-4
0	-67	8	-3	2	3	-2	11
0	-5	-1	2	0	2	1	-1
0	-5	-1	2	0	2	1	-1
-2	2	-6	7	0	1	1	1
0	7	-6	-9	-3	0	0	4
2	5	-3	-2	-3	0	0	2
-3	4	-1	-2	2	-6	-2	0
0	3	-8	9	-6	4	-5	-3
2	-7	6	6	-6	0	2	9
-4	-2	9	-3	3	4	2	0
5	2	-7	5	1	4	1	-5
-8	6	-13	-3	-4	-3	1	4
-6	0	0	9	0	3	-2	-2
5	6	1	-4	3	6	2	-8
4	-12	8	4	3	4	1	1
4	2	-5	3	-3	2	-5	-4
2	7	-3	-5	-4	-12	-2	-6
7	18	6	6	-2	22	-6	2
3	10	-10	0	-6	6	2	-2
3	5	-2	10	-10	2	2	7
-2	9	8	6	3	-10	0	-3
9	-12	7	-20	-4	5	2	3
-24	-10	4	-2	17	3	7	-5
-19	-10	1	8	-4	0	1	-1
4	-9	1	-6	3	6	2	-1
0	-9	-4	11	-3	4	4	8
-17	6	0	0	0	1	1	-6
5	9	-5	-1	3	5	-1	-1
10	6	12	12	0	4	-1	-7

Table 6.3: Rank differences between top 50 3-word pattern types

The reduction in frequency of occurrence of this pattern arises from the fact that the isotextual subcorpus contains fewer tokens than the isolexical version. As shown in Table 6.4 below, in the isolexical subcorpus *in this letter* occurs 680 times in *Electronics Letters*, 290 times in *Applied Physics Letters*, and once in the *Journal of Applied Physics*. In the isotextual subcorpus this pattern occurs 88 times in *Electronics Letters,* 60 times in *Applied Physics Letters* and once in the *Journal of Applied Physics*. Since this pattern refers to the text in which the pattern occurs (the genre of the research article in these journals is referred to as a "letter") it would seem logical that this pattern should occur less frequently in a corpus of 500 texts (the isotextual version (see Appendix 3.2)) than it would in a corpus of 1909 texts (the isolexical version), because if there are fewer texts in the corpus there are fewer instances of texts referring to themselves.

What is more surprising from these results, however, is that *in this letter* is the only 3-word pattern which has such a high rank difference. Table 6.3 shows that few patterns have a rank difference of more than 20, and all these are in the lower reaches of the frequency table. Another pattern on the frequency list, *in this paper,* has a similar form and contains a near synonym of *letter* in this context, and also refers to the text in which it occurs. The difference in frequency of its occurrence in the two subcorpora is, however, not reflected in its rank difference, as also shown in Table 6.4 below. Instead of a high negative rank difference, its rank difference of 18 is both smaller and positive: its isotextual rank (21st with 346 occurrences) is higher than its isolexical rank (39th with 701 occurrences). It again does not appear to be an idiosyncratic use: it occurs in 312 texts isolexically and in 150 texts in the isotextual subcorpus, and so is again not used repeatedly in a small number of texts.

	in this letter		*in this paper*	
	isolexical frequency	isotextual frequency	isolexical frequency	isotextual frequency
Electronics Letters	680	88	0	0
Applied Physics Letters	60	60	2	1
Journal of Applied Physics	1	1	15	0
EEE Transactions on Microwave Theory and Techniques	0	0	436	191
IEE Proceedings – Electric Power Applications	0	0	195	146

Table 6.4: Occurrences of *in this letter* and *in this paper*

This chapter now continues with the quantitative results for 4-word fixed collocational patterns.

6.4.3 Isolexical and isotextual frequency rankings for the top 50 4-word fixed collocational patterns in different disciplines

This section presents the frequencies of 4-word patterns in the corpus in each discipline. Figure 6.9 and Figure 6.10 below show a graphical representation of the isotextual and isolexical frequencies for the top 50 patterns, again giving an idea of the distribution of patterns in each discipline and showing the distributions for the disciplines in relation to each other. Figure 6.9 shows the differences between disciplines in the isolexical comparison, while Figure 6.10 shows the differences in the isotextual comparison. The distributions are also shown in tabular form in Appendices 12-19, which also show which pattern occurs in each rank in the distribution.

The isolexical comparison between the subcorpora shown in Figure 6.9 reveals that 4-word patterns occur most frequently in the four Science disciplines. Nuclear Physics contains the most frequent pattern, *as a function of,* which occurs 2092 times (see Appendix 15). The frequencies of the most frequent 4-word patterns in the top four disciplines are less spread out than those for the isolexical 3-word patterns described above. The discipline with the next most frequent 4-word pattern is Cell Biology, in which *in the presence of* occurs 1970 times (see Appendix 12), and then Electrical Engineering (*as a function of*, 1837 times; see Appendix 13), and Mechanical Engineering (*as a function of,* 1410 times; see Appendix 14).

The 4-word pattern frequency curve for the first four disciplines is initially very steep, as the most frequent pattern occurs many times more than the next pattern, which in turn occurs many times more frequently than the next pattern, and so on. After the first five patterns the difference between the frequencies of successive patterns in the list becomes gradually smaller, and so the frequency curve flattens rapidly. The isolexical frequency distributions for 4-word patterns again have a "long tail" shape, so after the first 30 patterns the frequency curves for all disciplines are again almost

identical in shape, with the difference in frequencies between them remaining largely constant.

After the four Science disciplines, the discipline with the next most frequent 4-word pattern is Marketing, in which *on the basis of* occurs 1024 times (see Appendix 17), followed by Philosophy (*in the case of,* 826 times; see Appendix 18), Sociology (*in the united states,* 767 times; see Appendix 19), and Economics (*in the united states,* 735 times; see Appendix 16). It can be seen from Figure 6.9 that the frequency distributions for these latter four disciplines are flatter, being less steep initially, and begin to converge earlier than for the first four disciplines. After 15 patterns the frequency curves for all disciplines are almost identical in shape.

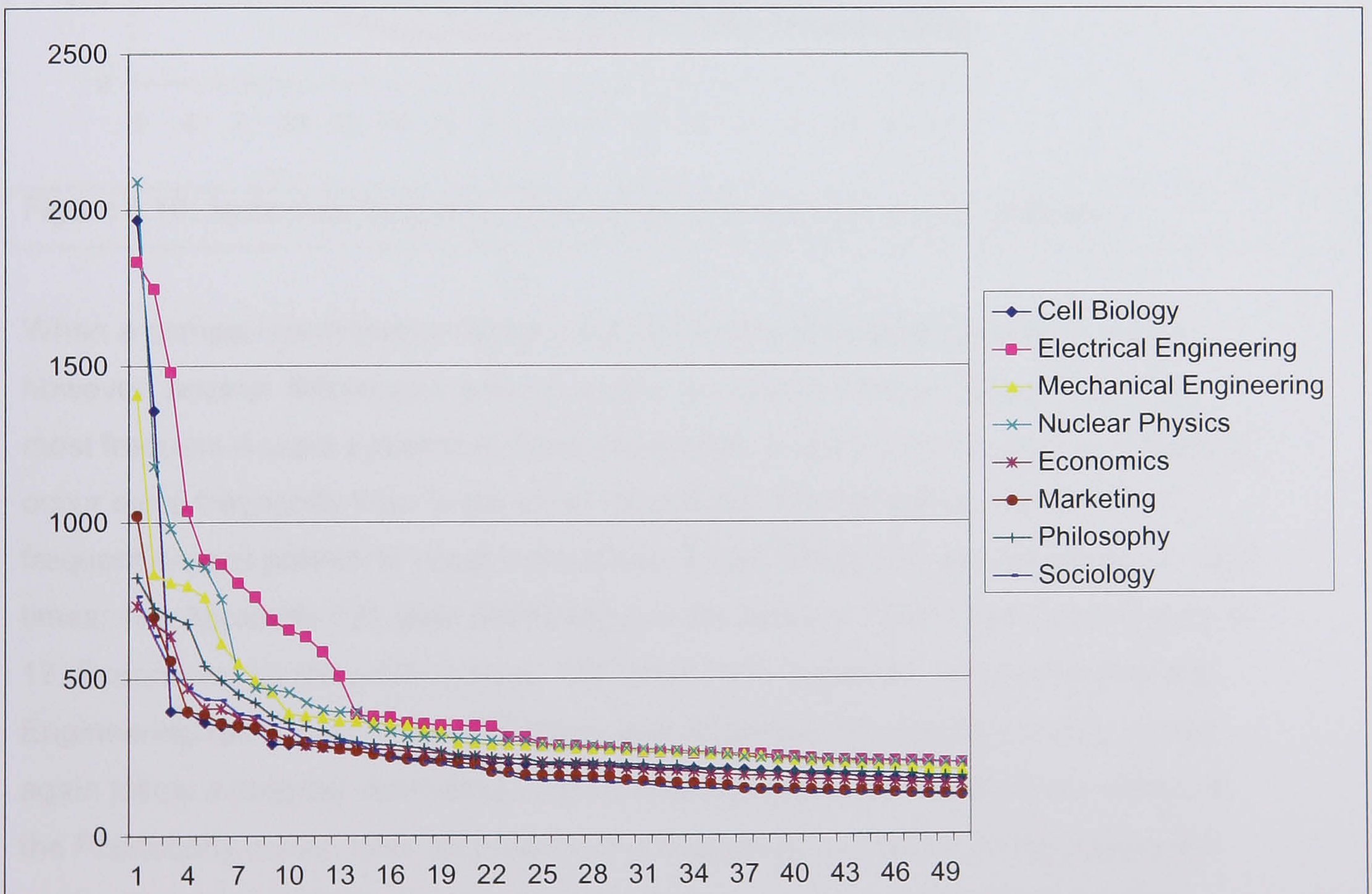


Figure 6.9: Isolexical frequency distribution of the top 50 4-word patterns

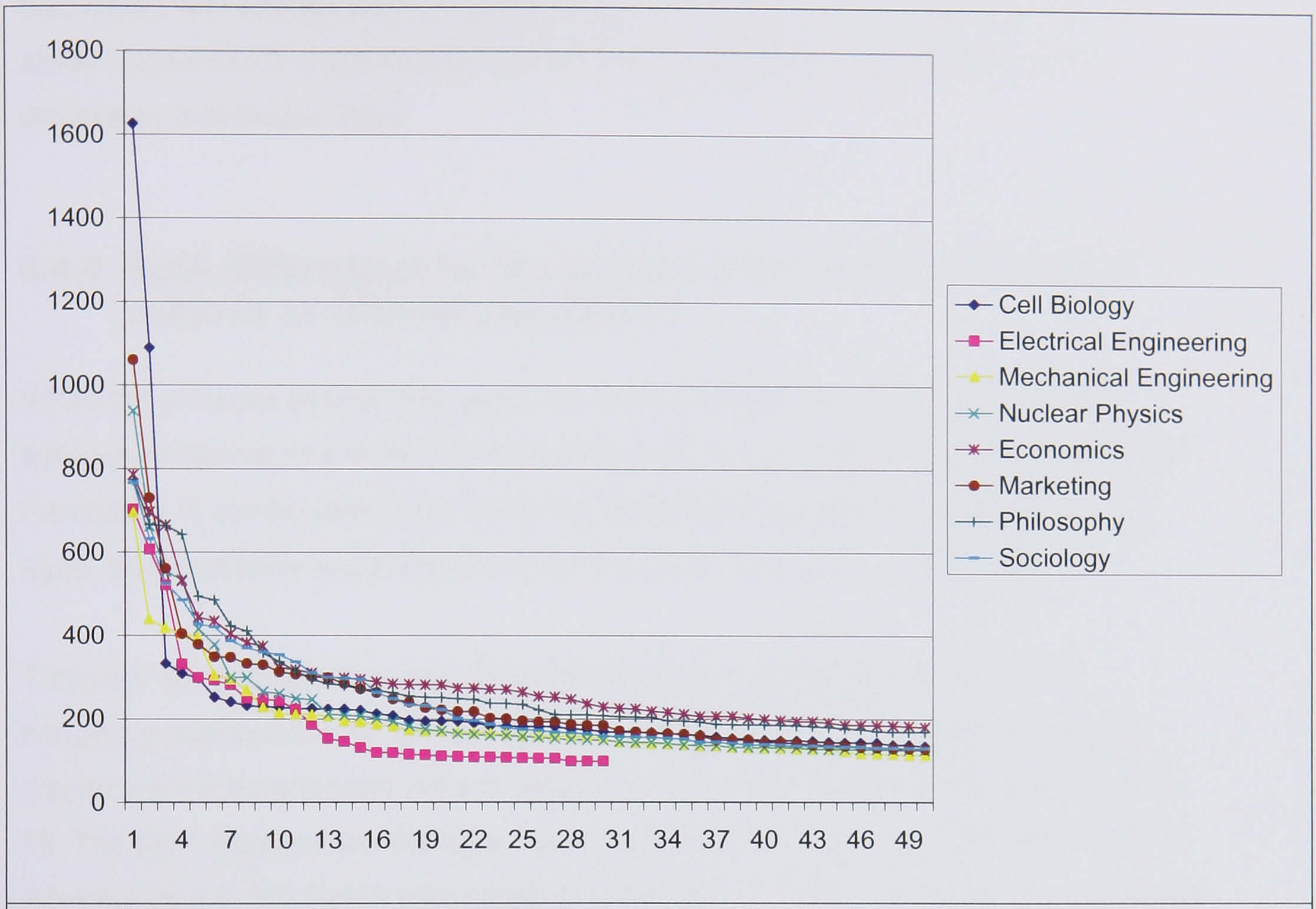


Figure 6.10: Isotextual frequency distribution of the top 50 4-word patterns

When a comparison between isotextual frequencies of 4-word patterns is made, however, several differences are noticeable, as can be seen in Figure 6.10. The most frequent 4-word patterns in three disciplines, in both Science and non-Science, occur more frequently than in the other disciplines. The discipline with the most frequent 4-word pattern to occur isotextually is Cell Biology (*in the presence of*, 1627 times; see Appendix 12), then Marketing (*on the basis of,* 1062 times; see Appendix 17, Economics (*in the united states*, 787 times; see Appendix 16) and Mechanical Engineering (*as shown in fig*, 696 times; see Appendix 14). The frequency curves again follow a long tail distribution, with a less noticeable convexity to the shape of the Philosophy curve. After 20 patterns the frequency curves for all disciplines are again almost identical in shape, with the difference in frequencies between them remaining largely constant.

There seems to be a weaker relationship between the length of texts in a discipline and the frequency of the most frequent 4-word patterns, compared to that for 3-word patterns. In the isolexical comparison the most frequent 4-word patterns in the

disciplines with shorter texts (and therefore more texts in their subcorpus) do not always occur much more frequently than the frequent 4-word patterns in the disciplines with longer texts.

6.4.4 Rank differences for the top 50 4-word fixed collocational patterns in different disciplines

While the previous section has again revealed that there is clearly a difference between isolexical and isotextual frequencies of 4-word patterns in each discipline in the corpus, it can be seen from Table 6.5 below that this difference in frequency again does not have much effect on the frequency rankings of these patterns.

Table 6.5 summarises the rank differences between isolexical and isotextual frequency distributions for the top 50 4-word patterns in each discipline. Its eight columns are the right-hand column taken from each of the tables in Appendices 12-19. The third column from the right, for example, contains the rank differences for 4-word bundles in Marketing calculated in Appendix 17. Putting the rank differences for all eight disciplines in one table again allows the rank differences in each discipline to be compared at a glance. It can be seen that in the top 10 rank positions in all subcorpora no patterns have a rank difference of more than 5 places. This is again regardless of the VSD: for example, the most isolexically frequent pattern in Electrical Engineering (0.34) is also the most isotextually frequent, i.e. they are both ranked first in the distribution; this gives a difference in rank for *shown in fig* of zero between the isolexical and isotextual versions.

The one rank difference to stand out from the others is for the pattern *is an element of*, which, as Appendix 16 shows, occurs in Economics far less often isotextually (ranked 54th with 159 occurrences) than it does isolexically (ranked 9th with 324 occurrences), and so has a rank difference of -45. This difference is the result of an idiosyncratic use which has not been excluded by the criterion requiring a fixed collocational pattern to occur in at least five texts. In the isolexical version the 324 occurrences come from 31 texts, which passes the criterion, but in a few texts in the *Economic Journal* (files 0020071090405, 406, 407, and 408) it is used many times.

Cell Biology (0.80)	Electrical Engineering (0.34)	Mechanical Engineering (0.52)	Nuclear Physics (0.54)	Economics (1.09)	Marketing (1.00)	Philosophy (0.94)	Social Science (1.05)
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
-2	0	0	0	0	0	0	0
1	-2	0	-2	0	0	0	0
1	-5	0	1	-1	0	0	-1
-2	1	0	1	1	-1	0	1
-5	3	0	0	0	-2	0	-3
1	1	0	0	0	2	0	1
0	-2	0	2	-45	1	0	1
0	2	-1	1	-2	0	0	-1
0	2	1	-1	0	-2	0	2
6	-	-3	-3	-2	-2	0	-3
-1	1	-1	2	3	-2	-3	1
1	-2	1	4	-4	-3	-1	1
0	-4	-1	-5	1	3	-8	1
3	2	6	3	-1	-3	2	-1
0	4	-6	-12	3	-1	4	1
1	-7	6	-5	2	-3	1	0
0	-6	-5	0	-3	3	1	0
0	5	-4	5	-5	-2	-1	-2
-2	4	-8	-6	2	1	2	0
-1	-	-5	-3	1	-1	2	2
2	3	0	-11	4	-1	1	-3
4	8	4	4	5	-1	-3	1
2	7	-12	-7	4	-1	-4	-2
4	-	2	4	1	-2	-1	-3
-3	3	-6	10	3	1	-1	3
6	7	7	-19	1	6	1	-7
2	-	9	-2	0	2	-3	1
-5	-	8	-20	-4	2	4	-16
7	-	12	0	3	2	-3	-1
7	9	7	9	-3	-4	-7	-2
-15	8	4	9	2	0	-4	2
6	-	-2	5	0	3	4	4
-2	-	-4	7	2	3	-12	-4
1	-	8	18	4	6	-6	-2
6	14	4	4	-1	2	1	-4
2	-	1	0	-2	-5	-3	-7
-2	-	5	13	2	0	5	5
10	-	3	2	-13	4	3	14
2	-	2	1	-4	4	-12	2
9	15	7	-1	6	1	15	-8
15	-	-5	-12	-6	6	2	0
10	-	12	-2	1	1	0	7
1	-	13	-2	3	5	-1	-9
-5	20	2	8	-10	2	6	-4
-3	-	5	4	-2	-1	0	-7
5	-	5	7	5	-1	-4	-5
9	-	3	-	-3	1	4	4
7	-	6	-9	-2	-4	1	7

Table 6.5: Rank differences between top 50 4-word pattern types

This is the result of the articles containing mathematical proofs which might use this pattern ten times in one sentence, as underlined in Example 6.1 below:

> Example 6.1: It is customary to write, for all x, y is an element of X, xRy if and only if (x, y) is an element of R. When a binary relation R satisfies completeness (For all x, y is an element of X, x not equal to y implies {(x, y) is an element of R or (y, x) is an element of R}), reflexivity (For all x is an element of X, (x, x) is an element of R) and transitivity (For all x, y, z is an element of X, {(x,y) is an element of R & (y, z) is an element of R} implies (x, z) is an element of R), we say that R is an ordering on X.

Purely as a result of the sampling methodology, those texts in which *is an element of* occurs frequently were not included in the isotextual version of the Economics subcorpus. The isotextual version needed to include the 100 most recent texts from each journal, and, as Appendix 3.5 shows, 60 texts were available from the *Economic Journal* in the most recent year, 2000, and so only 40 texts needed sampling from this journal for 1999. The most recent 40 texts from 1999 were those from 0020071090409 to 809 and so these were included, meaning that texts 405-408, which contained many of the occurrences of *is an element of*, were excluded. Had these four papers been published just one issue later, they would have been included in the isotextual corpus and the isotextual frequency of this pattern would have been higher. In fact, the isotextual frequency of occurrence of this pattern is much lower - 324 occurrences from 24 texts – than its isolexical frequency, and the pattern thus falls in the rankings, which leads to it having a higher rank difference.

Lower down the frequency table, rank differences become greater as the tail of the distribution is reached. At this point in the distribution patterns occur with very similar frequencies, so a small change in frequency of a pattern can have a larger effect on its ranking in the frequency table, making the rankings less stable overall.

It can also be seen in Table 6.5 that no rank difference is given for some patterns in some disciplines. This is because the isotextual frequency of these clusters is too low for them to be considered as fixed collocational patterns, and so their rankings cannot be compared.

This chapter continues with the quantitative results for the second research question, concerning the fixed collocational patterns occurring across disciplines.

6.5 Fixed collocational patterns occurring frequently across all disciplines

Having looked at the frequency rankings and structural forms of frequently occurring fixed collocational patterns, we now turn to identification of those patterns which occur in all eight disciplines. No 5, 6, or 7-word patterns were found to occur across all eight disciplines, and so this section will present only results for 3 and 4-word patterns.

6.5.1 3-word fixed collocational patterns occurring frequently across all disciplines

Overall, more 3-word patterns occur across non-Science disciplines than occur across Science disciplines. 166 3-word pattern types were found to occur isolexically in all four of the Science disciplines, while a total of 226 3-word pattern types were found to occur isolexically in all four of the non-Science disciplines (see Table 6.6 below). The difference was noticeably larger between the isotextual subcorpora. 64 3-word pattern types were found to occur isotextually in all four of the Science disciplines, while a total of 227 3-word pattern types, almost the same number as occurred isolexically, were found to occur isotextually in all four of the non-Science disciplines.

	3-word patterns	
	Isolexical	Isotextual
Science disciplines	166	64
Non-Science disciplines	226	227
All disciplines	77	42

Table 6.6: Totals for 3-word pattern types occurring isolexically and isotextually in Science, non-Science, and all disciplines

To a large extent the differences between isolexical and isotextual versions seem to reflect the version size differentials in the different disciplines. The non-Science disciplines each have a VSD of 1.0 or greater, and so similar numbers of 3-word pattern types occur in each version of the subcorpus. The non-Science disciplines have a VSD of less than 1.0, and so far fewer 3-word pattern types occur in each

version of the subcorpus. The version size differential does not, however, explain the overall difference between Science and non-Science disciplines. In both versions, non-Science disciplines share more 3-word patterns than Science disciplines, regardless of the number of texts or tokens.

There are many fewer 3-word fixed collocational patterns shared between Science and non-Science disciplines. A total of 77 3-word pattern types were found to occur across all eight disciplines in the isolexical version of each subcorpus, and a total of 42 3-word pattern types were found to occur across all eight disciplines in the isotextual version of each subcorpus. Considering that no discipline has fewer than 800 isolexical 3-word patterns, and in six out of the eight disciplines there are more than 600 isotextual 3-word pattern types (see Figure 6.1 above), the figures of 77 and 42 seem rather low. The isolexical subcorpus with the fewest 3-word pattern types, Sociology, has 867, and so only around 9% of its 3-word patterns occur in all the other disciplines. The percentage for the other subcorpora was even lower. The isotextual subcorpus with the fewest 3-word pattern types, Electrical Engineering, has 285, and so only 15% of its 3-word patterns occur in all the other disciplines. The percentage for the other subcorpora again was even lower.

In view of the relationship between frequent isolexical and frequent isotextual fixed collocational patterns identified above, it was decided to focus on those patterns which were frequent both isolexically and isotextually. Table 6.7 below shows these 42 patterns in the left hand column, and gives the isolexical and isotextual frequencies for each pattern. It also shows the rankings in the frequency distribution for each version of the corpus. There are gaps in the frequency rankings because some patterns which are frequent in the corpus as a whole do not occur more than 100 times in every discipline. For example, the most frequent 3-word pattern in the isolexical corpus as a whole is *shown in fig*, but since it does not occur across all eight disciplines, it is not included in the list. The second most frequent 3-word pattern, *in order to*, which does occur more than 100 times in each discipline, is therefore the first item in the list.

	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference	Ellis & Simpson rank
in order to	10964	2	7360	4	-2	2
the number of	10543	3	8228	1	2	4
as well as	10305	4	7963	3	1	1
due to the	8316	6	4729	15	-9	-
the case of	7650	7	5583	9	-2	13
with respect to	7507	8	5627	8	0	-
the fact that	7325	9	6080	5	4	-
the presence of	7235	11	5463	10	1	-
there is a	6859	12	5881	7	5	5
a function of	6809	13	3901	24	-11	-
one of the	6316	14	4853	14	0	-
based on the	6055	15	3862	26	-11	15
there is no	5896	16	5112	12	4	6
in the case	5782	17	4184	21	-4	17
part of the	5754	18	4373	17	1	-
the use of	5449	22	4197	20	2	3
in which the	5109	24	4301	19	5	18
in this paper	4645	28	3406	32	-4	-
the other hand	4625	29	3414	31	-2	12
a number of	4311	32	3597	27	5	7
as a result	3806	41	3038	36	5	14
according to the	3696	43	2586	54	-11	-
of the two	3664	44	2453	60	-16	-
analysis of the	3587	45	2677	49	-4	-
is that the	3516	46	2967	41	5	-
because of the	3489	47	2481	57	-10	-
the results of	3424	51	2458	58	-7	-
at the same	3124	60	2427	62	-2	-
in the present	2976	65	1945	99	-34	-
in the same	2925	68	2374	66	2	-
that of the	2882	71	1664	133	-62	-
such as the	2846	74	2292	70	4	-
it is possible	2776	80	2157	76	4	-
it has been	2565	89	1676	130	-41	-
well as the	2520	94	1827	109	-15	-
between the two	2517	95	1843	107	-12	-
as in the	2368	112	1918	101	11	-
most of the	2357	113	1843	107	6	-
and in the	2251	123	1726	123	0	-
a result of	2027	146	1600	143	3	-
the analysis of	2026	147	1478	163	-16	-
to that of	1757	198	1052	274	-76	-

Table 6.7: 3-word patterns occurring across all disciplines in this study, compared with the academic writing part of the BNC and the most frequent written academic formulas of Ellis and Simpson (2005)

The rank differences are also shown in Table 6.7 above for each pattern in each version of the subcorpus. At first glance, it appears that the rank differences between isolexical and isotextual versions of the full corpus are less stable than the frequency rankings for individual disciplines described in the previous sections. Several patterns have rank differences higher than 30 places, and few have a rank difference of zero. However, only 12 patterns have a rank difference greater than 10 places. The top three combinations in the isolexical version, namely *in order to*, *the number of*, and *as well as*, are the same top three combinations as in the isotextual version.

These fixed collocational patterns can be compared with the written academic formulas of Ellis and Simpson (2005). The top 20 most significant written academic formulas consist of 17 3-word and 3 4-word fixed collocational patterns. 13 of Ellis and Simpson's 3-word academic formulas are also 3-word fixed collocational patterns which occur 100 times in all eight disciplines in the present study. The rankings of the same 3-word patterns in the study by Ellis and Simpson are shown in the right-hand column of Table 6.7. Despite the vastly different corpora collected, software employed, and identification procedures used in this study and in that of Ellis and Simpson, the similarities between the two sets of results are remarkable. The overlap between the two sets of results suggest that these 3-word patterns are in some way universal, regardless of the corpus data or methodology used.

6.5.2 4-word fixed collocational patterns occurring frequently across all disciplines

The tendency for greater numbers of patterns to be shared within non-Science disciplines than within Science disciplines was repeated in the case of 4-word patterns. In the isolexical version of each subcorpus a total of 6 4-word pattern types were found to occur in Science disciplines, while a total of 17 3-word pattern types were found to occur in non-Science disciplines (see Table 6.8 below). The difference was slightly larger between the isotextual subcorpora. 4 4-word pattern types were found to occur isotextually in all four of the Science disciplines, while a total of 17 4-word pattern types, again the same number as occurred isolexically, were found to occur isotextually in all four of the non-Science disciplines.

	4-word patterns	
	Isolexical	Isotextual
Science disciplines	6	4
Non-Science disciplines	17	17
All disciplines	3	3

Table 6.8: Totals for 4- word pattern types occurring isolexically and isotextually in Science, non-Science, and all disciplines

The differences between isolexical and isotextual versions again seem to reflect the version size differentials in the different disciplines. The non-Science disciplines each have a VSD of 1.0 or greater, and so similar numbers of 4-word pattern types occur in each version of the subcorpus. The non-Science disciplines have a VSD of less than 1.0, and so far fewer 4-word pattern types occur in each version of the subcorpus. The version size differential, however, again does not explain the overall difference between Science and non-Science disciplines. In both versions, non-Science disciplines share more 4-word patterns than Science disciplines, regardless of the number of texts or tokens.

There are many fewer 4-word fixed collocational patterns shared between Science and non-Science disciplines. The number of 4-word patterns occurring more than 100 times in every discipline is far lower than for 3-word patterns. Only three 4-word pattern types: *as a result of*, *as well as the*, and *in the case of* were found to occur across all eight disciplines in the isolexical version of each subcorpus, and three 4-word pattern types: *as well as the*, *in the case of*, and *on the other hand*, *were* found to occur across all eight disciplines in the isotextual version of each subcorpus. Considering that no discipline has fewer than 74 isolexical 4-word patterns, and no discipline has fewer than 40 isotextual 4-word pattern types (see Figure 6.1 above), the figures of 3 in both versions again seems rather low, and are lower in percentage terms than the results for 3-word patterns. The isolexical subcorpus with the fewest 4-word pattern types, Sociology, has 74, and so only around 4% of its 4-word patterns occur in all the other disciplines. The percentage for the other subcorpora is even lower. The isotextual subcorpus with the fewest 4-word pattern types, Electrical Engineering, has 30, and so only 10% of its 4-word patterns occur in all the other disciplines. The percentage for the other subcorpora again is even lower.

Only two 4-word pattern types occur across all eight disciplines in both the isolexical and isotextual versions of the corpus, as shown in Table 6.9 below. One of these, *in the case of*, is also one of Ellis and Simpson's 3 most significant 4-word academic formulas. This is fairly remarkable, given how few 4-word patterns occur in all disciplines in the research article corpus.

	Iso-lexical freq.	**Iso-lexical rank**	**Iso-textual freq.**	**Iso-textual rank**	**Rank diff.**	**Ellis & Simpson rank**
in the case of	10964	2	7360	4	-2	11
as well as the	10543	3	8228	1	2	-

Table 6.9: 4-word patterns occurring across all disciplines in this study, compared with the academic writing part of the BNC and the most frequent written academic formulas of Ellis and Simpson (2005)

6.5.3 Summary

These results show that fixed collocational patterns which are most frequent isolexically are also most frequent isotextually This relationship has been observed with five million words in each isolexical subcorpus and 500 texts in each isotextual subcorpus. Further comparison of different numbers of texts and tokens might establish if there is a minimum threshold below which this relationship no longer applies.

In order to investigate whether there is any similarity between the frequent *isotextual* patterns found in this study and frequent *isolexical* fixed collocational patterns from other studies, i.e. lexical bundles, the results from the present study can be compared with the frequent lexical bundles identified in Biber and Hyland. If frequent *isolexical* fixed collocational patterns which are frequent in genre corpora (i.e. the lexical bundles in Biber 2006) are found to be similar to *isolexical* fixed collocational patterns frequent in a register corpus (i.e. the lexical bundles in Hyland 2008), which are in turn found to be similar to *isotextual* fixed collocational patterns frequent in a genre corpus (i.e. the patterns identified in this study) then this would be more evidence of stability in the forms of frequent fixed collocational patterns. It is important, as Stubbs (2001: 124) points out, to check the findings from a corpus study "on different, *but comparable*, data, in order to see whether they were an

artefact of one single data set" (emphasis added). While a textbook genre corpus (Biber 2006) and an academic register corpus (Hyland 2008) do not immediately appear directly comparable, the point here is that we are not investigating the comparability of the corpora, but investigating the extent of the similarity between isotextual and isolexical fixed collocational patterns in different data sets.

In regard to the function of these fixed collocational patterns, the comparison of frequent isotextual fixed collocational patterns in this study with Biber (2006) and Hyland (2008) will also give a clearer idea of which patterns are more important in research articles in different disciplines, since they are by definition frequent across the same number of communicative acts, and are not biased by the imbalances between the subcorpora in which they were found. We now, therefore, move on to the comparison of the isotextual fixed collocational patterns in this study with the lexical bundles of Biber (2006) and Hyland (2008).

Chapter 7 Fixed collocational patterns and lexical bundles

7.1 Comparison of fixed collocational patterns with the lexical bundles of Biber (2006) and Hyland (2008)

This section attempts to compare the frequently occurring fixed collocational patterns in this study to the two most recent functional classifications of lexical bundles available, those of Biber (2006) and Hyland (2008). Biber (2006: 166-168) presents a genre-based study of the functions of lexical bundles found in the textbook section of the T2K-SWAL corpus, comprising the disciplines of Business, Engineering, Natural Science, Social Science, and Humanities disciplines. Hyland (2008: 13-14), on the other hand, presents a register-based study of the functions of lexical bundles in academic prose, from a corpus of research articles, dissertations and MA theses in the disciplines of Biology, Electrical Engineering, Applied Linguistics, and Business Studies.

These previous sets of results are not ideal for comparison as they contain differing amounts of detail and are not comprehensive lists of how many bundles occur with each function. Biber lists "common" (2006: 166) lexical bundles and uses asterisks to indicate their frequency (see the left of Table 7.1 below), but he is not specific about whether the bundles in his results comprise all the bundles with these functions, or merely the frequent ones. In specifying his functional framework (see the right of Table 7.1 below), Hyland gives two or three examples of bundles with each function, but again does not list all the bundles which were identified as having that function.

The frequent isotextual fixed collocational patterns in the present study can nonetheless be matched with those lexical bundles found in Biber's and Hyland's studies. It can be assumed that both Biber and Hyland chose to include their examples in their results because these were the most salient or typical. It will therefore be relevant to find whether frequent bundles found to have particular functions by Biber and Hyland are also frequent isotextually in the present study. It will also be important to find out whether any lexical bundles are "missing" from their results as a consequence of their decision to assign genre-related textual discourse

functions to lexical bundles found in register corpora. A summary of the two functional taxonomies is given in Table 7.1 below:

Key to Biber's lexical bundle functions:	Key to Hyland's lexical bundle functions
Stance bundles	**Research-oriented**
1 Epistemic stance - impersonal	1 Location
2 Attitudinal/Modality stance	2 Procedure
a) Ability/effort: Impersonal	3 Quantification
b) Importance: Impersonal	4 Procedure
	5 Topic
Discourse organizers	
	Text-oriented
Referential bundles	1 Transition signals
1 Identification	2 Resultative signals
2 Specification of attributes	3 Structuring signals
a) Quantity/mathematical expression	4 Framing signals
b) Predicative	
c) Tangible framing attributes	**Participant-oriented**
d) Intangible framing attributes	1 Stance features
3 Time/place/text/other reference	2 Engagement
a) Place or institution reference	
b) General location reference or framing	
c) Text deixis	

Key to frequency symbols for Biber (2006) Table 7.2 to Table 7.7 below:	Key to frequency symbols for Hyland (2008) in Tables 23-27 below:	Key to isotextual frequency symbols for the present study in Tables 23-27 below:
* 20-39 occurrences per million words	*** occurs in the top 50 lexical bundles	* 0-199 occurrences
** 40-99 occurrences per million words		** 200-499 occurrences
*** >100 occurrences per million words		*** >500 occurrences

Table 7.1: Summary of Biber's (2006) and Hyland's (2008) functional taxonomies for lexical bundles; keys to symbols used in Tables 7.2-7.7

Table 7.1 also provides the key to each type of function used in the next six tables, from Table 7.2 to Table 7.7 in section 7.1.2, and which show the comparison between Biber's study, Hyland's study, and the present study.

7.1.1 Comparison of isotextual fixed collocational patterns with lexical bundles: a brief example

An initial example is given here of the findings presented in the following six tables. One left-hand column and one right-hand column of a table display all the available information regarding Biber's and Hyland's lexical bundles respectively, while the table's central columns show the isotextual fixed collocational patterns found in the

this study. To the left of each table a column shows Biber's functional categories and subcategories, and lists the lexical bundles which he identifies in his study as performing that function in textbooks. Asterisks are used to indicate how common that lexical bundle is in his corpus (the key is given in Table 7.1).

The right-hand column of each table lists the function which Hyland assigns to the same lexical bundle in that row, if the information is available. It also gives indication of whether it occurs in the top 50 lexical bundles in any discipline in his corpus. For a small number of lexical bundles, such as *the operation of the* in Table 7.6, there is no such indication. This is because while Hyland lists such bundles as having particular functions in his specified framework, they are not in fact listed in his tables showing the top 50 bundles in each discipline. A cross † is used to denote such bundles. While Hyland's framework is for 4-word lexical bundles, he also, presumably in error, lists a 5-word bundle *it should be noted that* as an example of a bundle with a "participant-oriented engagement" function.

The central columns in the tables indicate whether isotextual fixed collocational patterns with the same form as these lexical bundles occur frequently in any disciplines in the present study. A summary of the content of each table is given in the two columns to the far left. The first shows the total number of disciplines the pattern occurs in, and the second shows the number of asterisks for each pattern in total. The more asterisks in the row for any lexical bundle, then the more overlap there is between the three corpora: a count of 1-15 is considered low, 16-30 medium, and more than 31 is considered high.

An example of the insights which can be gained from this comparison is given by the lexical bundle *at the same time,* which is shown near the bottom of Table 7.2. In Biber's framework this lexical bundle functions as a "discourse organizer" (left-hand column) while in Hyland's framework (right-hand column) it is assigned the function of a "research-oriented location signal," shown by the abbreviation "Research 1." It can be seen in the table that *at the same time* occurs as a frequent lexical bundle in every discipline in both Biber and Hyland, and also frequently in non-Science disciplines in my research article corpus, occurring in a total of 13 disciplines with 32 asterisks. This finding reveals a number of things:

The lexical bundle *at the same time* is highly frequent in the wider register of academic writing, as it occurs in every discipline in Hyland (2008). It is also frequent in texts belonging to a different genre, that of academic textbooks, since it occurs in every discipline in Biber (2006). It is occurs both isolexically and isotextually frequently in non-Science disciplines in my research article corpus, and so it is a frequent fixed collocational pattern which occurs independently of text length. If this bundle is also assigned a function by both Biber and by Hyland, it then becomes a candidate for further detailed study. The next section summarises the findings from Table 7.2 to Table 7.7.

7.1.2 Comparison of isotextual fixed collocational patterns with lexical bundles: full results

The large numbers of gaps in tables 23-28 reveal, however, that there is not much overlap between the results given by Biber and those given by Hyland. Only four bundles are assigned functions by both Biber and Hyland in Table 7.2, three in Table 7.3, five in Table 7.4, and two in Table 7.5. The functions and lexical bundles in Table 7.6 are those of Hyland's framework only, as there was no overlap between these and Biber. This lack of overlap reveals that there is a limited number of 4-word patterns which are common across academic textbooks and the register of academic prose. Those patterns which *are* common therefore, such as *on the other hand*, consequently assume more importance in this study.

Table 7.2 below shows the lexical bundles identified by Biber (2006) as being in the first two of his three main functional categories: stance bundles and discourse organizers. No stance bundles occur frequently across all the studies. The most frequent text organising bundles are *on the other hand, as well as the,* and *at the same time.* The text-organising transition function assigned by Hyland to *on the other hand* is in agreement with Biber's, while Hyland does not assign a function to *as well as the*. The text-organising transition function assigned by Hyland to *at the same time* is instead classed as "research oriented location." This kind of bundle for Hyland "help[s] writers to structure their activities and experiences of the real world" (Hyland 2008: 13) such as the time in which an event is located.

Table 7.3 below shows the lexical bundles identified by Biber (2006) as being in the second main functional category of referential bundles. There is very little overlap between the studies for this function. Only one bundle, *as a function of* occurs with even medium frequency, and this is mainly in Science disciplines, although it is also found in Economics and Marketing research articles in the present study.

Table 7.4 below shows lexical bundles in more of Biber's referential subcategories. Six of these occur with medium frequency. Three of these, *in terms of the, in the form of,* and *in/of the united states/us* are not assigned a function by Hyland, however. The bundle *the size of the*, which occurs almost equally in Science and non-Science disciplines in all three studies, is assigned by Biber a "tangible framing attribute" function, while Hyland assigns it another "research-oriented location" function. *As a result of* for Hyland is a "text-oriented resultative signal", while for Biber it is an "intangible framing attribute". The functions assigned by Biber and Hyland are more in agreement for *on the basis of,* which occurs mostly in non-Science disciplines as both their functions are topic-related: Hyland assigns it a "research-oriented topic" function while for Biber it has an "intangible framing attribute" function.

The most frequent bundle in Table 7.4 is *in the case of*, which occurs in every discipline in every study, one of the only bundles to do so. It is also the only 4-word bundle which occurs in every discipline in every study and is also found in Ellis and Simpson's top 20 written academic formulas. The functions assigned in the two studies are also fairly similar: Biber assigns *in the case of* an "intangible framing attribute" function, while Hyland terms it a "text-oriented framing signal." Its ubiquity in all these studies suggests that *in the case of* is worthy of further investigation, and so a case study will be made of it in the next chapter.

Only one bundle in Table 7.5 deserves mention: *as shown in figure/fig*. This is common in Science disciplines, and is assigned similar functions by Biber and Hyland: "referential text deixis" and "text-oriented structuring signal" respectively. Table 7.6 lists those bundles which are not mentioned by Biber but are seen as important by Hyland. The one bundle of interest here is *with respect to the,* termed a

"text-oriented framing signal" by Hyland. This occurs in most disciplines both in Hyland's study and the present study.

The final table in the group is Table 7.7, which contains fixed collocational patterns rather than lexical bundles. All these patterns occur frequently in more than one discipline in the present study, but do not occur in Biber's or Hyland's study. They do not occur in the genre of textbooks but they might have if Biber had used a large enough corpus for an isolexical study to be made without the need to normalise results. Biber's textbook subcorpora ranged from 72,000 tokens for Engineering to 213,000 for Social Science, which meant that all results had to be multiplied to reach the isolexical definition of a lexical bundle of occurrences per million words. These patterns might also have been found in Biber's study if he had used a isotextual corpus which contained the same number of each textbooks in each discipline so that an equal number of communicative events were included. The number of texts in his corpus ranged from 9 in the Engineering subcorpus to 21 in the Social Science subcorpus (see Figure 4.10 above). Similarly these patterns might have occurred in Hyland's study if he had used a large enough corpus for an isolexical study to be made without the need for normalising results. Only one of his subcorpora had more than a million tokens, while the rest ranged from 632,500 in the case of the Electrical Engineering subcorpus to 844,400 in Business Studies (see Figure 4.12 above). As his register corpus contains an unknown proportion of texts in each genre, it is difficult to see which parts of the discourse outweigh the others.

These patterns in Table 7.7, if they are seen as lexical bundles, could be called the "missing" bundles. They are missing because of the isolexical methodology for identifying lexical bundles: they are missing from Hyland's and Biber's results presumably because they are not frequent enough in isolexical terms to meet their isolexical identification criteria. These fixed collocational patterns, however, are isotextually frequent and so occur frequently irrespective to the length of texts or number of tokens in the corpus.

The tables discussed in this section follow over the next six pages.

			Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
		4-word pattern/bundle	**Bus.**	**Eng.**	**Nat. Sci.**	**Soc. Sci.**	**Hum**	**Cell Bio.**	**Elec. Eng.**	**Mech. Eng.**	**Nucl. Phys.**	**Econ.**	**Mktg.**	**Phil**	**Soc**	**Bio.**	**Elec. Eng**	**App. Ling.**	**Bus. Stud.**	**Hyland 's function**
Disc.	***s**	**Stance 1**																		
7	15	are more likely to	**			**						*	**		**			***	***	
3	4	are likely to be†	*			*							**							Participant 1
2	3	by the fact that				**	*													
10	19	the fact that the		**		**	*			*	**	**		**	*			***	***	
		Stance 2a																		
9	18	can be used to	**	***	**	*			*	*			**				***		***	
2	3	it is difficult to	*	**																
10	15	it is possible to	**	**			*		*	*	*	*		**	*		***			
1	2	can be calculated from		**																
2	3	can be determined by		**	*															
		Stance 2b																		
10	18	it is important to	**	**	*	*	*						**	*	**			***	***	
3	5	an important role in			*	*										***				
2	2	of the most important			*	*														
		Discourse organizers																		
17	40	on the other hand	**	***	***	**	**	*	**	**	***	***	**	***	***	***	***	***	***	Text 1
6	12	on the one hand	*			**	**							**	**			***		
17	34	as well as the	**	**	**	**	**	*	*	*	**	**	**	*	**	***	***	***	***	
2	2	for the most part	*				*													
1	2	we can think of	**																	
4	7	as long as the		**	*	*												***		
6	11	in this case the		**					*	*	**	**					***			
1	2	tells us that the			**															
13	32	at the same time	**	**	***	***	*					**	**	**	***	***	***	***	***	Research 1
6	11	in addition to the			**	*					*		*			***			***	Text 1

† Hyland lists this in his framework as having this function (2008: 14) but does not list it as being in the top 50 in any discipline in his tables (ibid.: 12)

Table 7.2: Comparison of Biber's (2006) lexical bundle functions (Stance and Discourse Organizers) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study

			Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
		4-word pattern/bundle	Bus.	Eng.	Nat. Sci.	Soc. Sci.	Hum	Cell Bio.	Elec. Eng.	Mech. Eng.	Nucl. Phys.	Econ.	Mktg.	Phil	Soc	Bio.	Elec. Eng	App. Ling.	Bus. Stud.	Hyland 's function
Disc.	*s	**Referential 1**																		
6	13	is one of the			**	*									*	***	***	***		
2	3	is referred to as	*	**																
6	15	one of the most		**	***	**	***								**			***		Research 3
1	2	is known as the			**															
		Referential 2a																		
5	8	the rest of the	*	**	*	**						**								
3	6	a great deal of	**	**		**														
5	6	a large number of	**		*		*					*	*							
7	15	a wide range of	**			**							*		*	***		***	***	Research: 3
4	6	the sum of the	**		**						*	*								
1	3	degrees of freedom of		***																
2	3	an increase in the		**	*															
3	6	of the number of		**	*														***	
9	9	the magnitude of the		***	**					*							***			Research 3
2	5	the rate at which		***	**															
3	5	the ratio of the		**	**						*									
9	22	as a function of		***	***	*			***	**	***	**	**				***			
		Referential 2b																		
4	6	is equal to the	*	**	**							*								
3	5	is given by the		**	*						**									
2	4	is proportional to the		**	**															

Table 7.3: Comparison of Biber's (2006) lexical bundle functions (Referential 1, 2a and 2b) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study

			Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
		4-word pattern/bundle	**Bus.**	**Eng.**	**Nat. Sci.**	**Soc. Sci.**	**Hum**	**Cell Bio.**	**Elec. Eng.**	**Mech. Eng.**	**Nucl. Phys.**	**Econ.**	**Mktg.**	**Phil**	**Soc**	**Bio.**	**Elec. Eng**	**App. Ling.**	**Bus. Stud.**	**Hyland 's function**
Disc.	***s**	**Referential 2c**																		
9	17	the size of the			**	**			*	*	*	***	*				***		***	Research 4
11	22	in the form of	**	**	**	**	*			*			*		**		***	***	***	
1	3	in place of the		***																
1	3	the composition of the		***																
3	6	the length of the		**						*							***			
2	3	the mass of the			**						*									
		Referential 2d																		
5	13	the nature of the											**		**	***		***	***	
17	39	in the case of	*	***	**	**	**	*	**	**	***	**	**	***	**	***	***	***	***	Text 4
13	26	in terms of the	*	**	**	*	*			**	***	**	**	**	**			***	***	
10	24	as a result of	**		***	**	**					**	**		**	***		***	***	Text 2
6	12	in the absence of						***				**	**	*	*	***				
3	6	the context of the	**			*													***	
1	2	beyond the scope of		**																
1	3	the order in which		***																
1	3	the temperature of the		***																
1	3	the velocity of the		***																
11	25	on the basis of			***	**	**	*		*		**	***	***	**			***	***	Text 4
3	5	the basis of the			**	*							**							
3	6	the structure of the			*		**										***			Research 4
3	4	in the course of				*	**								*					
		Referential 3a																		
6	9	the united states and	**		*	**	*					**			*					
8	22	in/of the united states/us	***		***	***	**					***	**		***				***	

Table 7.4: Comparison of Biber's (2006) lexical bundle functions (Referential 2c, 2d and 3a) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study

			Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
		4-word pattern/bundle	**Bus.**	**Eng.**	**Nat. Sci.**	**Soc. Sci.**	**Hum**	**Cell Bio.**	**Elec. Eng.**	**Mech. Eng.**	**Nucl. Phys.**	**Econ.**	**Mktg.**	**Phil**	**Soc**	**Bio.**	**Elec. Eng**	**App. Ling.**	**Bus. Stud.**	**Hyland 's function**
Disc.	*s	**Referential 3b**																		
1	2	in the same direction		**																
1	2	at the base of			**															
1	2	at the bottom of			**															
1	2	in the vicinity of			**															
1	2	on either side of			**															
2	4	the position of the			***					*										
2	4	the surface of the†			***					*										Research 4
2	4	the top of the			***	*														
		Referential 3c																		
9	21	as shown in figure/fig		***	***	*		**	**	**	**					***	***			Text 3
3	5	as we have seen				**	*							**						
2	3	in our discussion of	*		**															
2	3	in this chapter we	*	**																
		Multifunctional reference																		
10	23	the end of the	***		**	**	**			*		**			**	***		***	***	
6	11	the beginning of the	*		*	*	**									***		***		
11	24	at the end of	***	**	**	**	**			*		**			*	***		***	***	

† Hyland lists this in his framework as having this function (2008: 13-14) but does not list it as being in the top 50 in any discipline in his tables (ibid.: 12)

Table 7.5: Comparison of Biber's (2006) lexical bundle functions (Referential 3b, 3c, and Multifunctional reference) with Hyland's (2008) lexical bundle functions and the fixed collocational patterns in the present study

			Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
Disc.	*s	4-word pattern/bundle	Bus.	Eng	Nat. Sci.	Soc Sci.	Hum	Cell Bio.	Elec. Eng.	Mech. Eng.	Nucl. Phys.	Econ.	Mktg.	Phil	Soc	Bio.	Elec. Eng	App. Ling.	Bus. Stud.	Hyland 's function
3	8	at the beginning of										**				***		***		Research 1
3	7	in the present study								*						***		***		Research 1 / Text 3
1	3	the use of the																***		Research 2
3	4	the role of the													*			***		Research 2
1	3	the purpose of the																***		Research 2
		the operation of the†																		Research 2
1	3	in the Hong Kong																	***	Research 5
		the currency board system †																		Research 5
		in contrast to the†																		Text 1
2	6	it was found that														***	***			Text 2
1	2	these results suggest that						**												Text 2
1	2	in the next section											*							Text 3
9	21	with respect to the							**	**	**	**	**	**		***	***		***	Text 4
6	11	in the presence of						*			*	**	*			***			***	Text 4
		with the exception of†																		Text 4
1	3	may be due to														***				Participant 1
3	6	it is possible that						**					*			***				Participant 1
2	4	it should be noted that††								*								***		Participant 2
2	2	as can be seen							*		*									Participant 2

† Hyland lists this in his framework as having this function (2008: 13-14) but does not list it as being in the top 50 in any discipline in his tables (ibid.: 12)

†† This bundle contains five words, apparently an error in Hyland

Table 7.6: Comparison of Hyland's (2008) lexical bundle functions with Biber's (2006) lexical bundle functions and the fixed collocational patterns in the present study

		4-word pattern/bundle	Biber (2006): common lexical bundles in textbooks					The present study: top 50 fixed collocational patterns in RAs (isotextual frequency)								Hyland (2008): top 50 lexical bundles in RAs, dissertations and MA theses				
		Biber's function†††	Bus.	Eng	Nat. Sci.	Soc Sci.	Hum	Cell Bio.	Elec. Eng.	Mech. Eng.	Nucl. Phys.	Econ.	Mktg.	Phil	Soc	Bio.	Elec. Eng	App. Ling.	Bus. Stud.	Hyland 's function†††
Disc.	*s	**Stance 1**																		
2	3	is likely to be										*	**							Participant 1
2	3	more likely to be											*		**					Participant 1
2	5	if and only if										***		**						Participant 1
		Referential 2c																		
3	5	the value of the								*	**	**								Research 4
3	5	to the extent that										**	*		**					Research 4
2	2	at the time of											*		*					
		Referential 2d																		
4	5	the case of the									**		*	*	*					Text 4
2	2	for the case of								*	*									Text 4
2	3	the way in which												**	*					Research 2
		Referential 3c																		
2	2	can be expressed as							*	*										Participant 2
3	4	can be written as								*	**	*								Participant 2
2	4	in this paper we									**	**								Text 3
2	3	in this section we									*	**								Text 3
3	5	shown in fig 1							**	**	*									Text 3
3	5	shown in fig 2							**	**	*									Text 3
3	5	shown in fig 3							**	*	*									Text 3
2	2	shown in fig 4							*	*										Text 3
2	2	shown in fig 5							*	*										Text 3
2	2	shown in fig 6							*	*										Text 3

††† Apparent closest matching function

Table 7.7: The "missing" fixed collocational patterns: those occurring isotextually in more than one discipline in the present study but which are not listed among Biber's or Hyland's lexical bundles

Having compared the frequencies of isotextual fixed collocational patterns in the present study with those of Biber and Hyland, this section now turns to describing their structural forms using the categories developed by Biber et al. (1999) for lexical bundles and adapted by Hyland (2008). While lexical bundles are purely an isolexically-defined construct, it has been shown above that they are very similar in form to isolexical and isotextual fixed collocational patterns, and so the same structural categories can be used for both.

7.2 Forms of 4-word fixed collocational patterns

Biber et al. needed to develop their descriptive formal categories because, since lexical bundles and fixed collocational patterns are defined empirically, their forms, like n-grams, do not necessarily map onto traditional ways of describing language. For most patterns, there is no simple one-to-one correlation between the words in the pattern and recognised structural categories. Biber et al. (1999: 999) found that "there are almost no lexical bundles representing complete structural units." Instead, most bundles span two structures and end in a function word such as an article or preposition. Thus *the end of the* consists of a noun phrase and the beginning of a prepositional phrase. The category for this form would be *noun phrase + of-fragment*. Similarly, in the case of *that it is not,* Biber et al. categorise it as *(verb phrase +) that clause fragment,* since it contains part of a verb phrase (*it is not*), and the *that* comes from the beginning of an extraposed *that* clause such as *it can be seen that*.

In a small number, less than 5% (Biber, Conrad, & Cortes 2004: 377) of cases, lexical bundles have existing structural correlates. For example the 4-word pattern *in the united kingdom* can be described as a prepositional phrase. Biber et al. (1999: 999) make the interesting observation that "when a lexical bundle is structurally complete in academic prose, it is typically a prepositional phrase that functions as a discourse signalling device," such as *in the next chapter* or *in the present study*. It is argued in this study that the isolexical nature of lexical bundles makes it difficult to say whether bundles *typically* have such discourse signalling functions. Since Biber's lexical bundles were compared between registers (in the case of Biber et al., conversation and academic prose), his two corpora were balanced for tokens, and

so they contained different numbers of texts and different numbers of communicative acts, and also contained text fragments and thus excluded an unknown number of acts within the communication. They therefore could not be sure that the full range of discourse functions for texts in those registers had been included in his data, and so any frequent functions observed in their data were not necessarily typical of texts in those registers.

Biber et al. only describe structural categories for 4-word lexical bundles, and so only the forms of 4-word fixed collocational patterns can be compared. It is not possible to compare quantitatively the forms of patterns with the lexical bundles of Biber et al. since, as it is a study of register, they give quantitative results about the relative numbers of different forms in each register: “most lexical bundles in conversation are building blocks for verbal and clausal structural units, while most lexical bundles are building blocks for extended noun phrases or prepositional phrases” (1999: 992). Hyland (2008: 10), however, does give quantitative data about the relative frequencies of the different structural forms of lexical bundles in different disciplines, namely between journal articles in two Science (Biology and Electrical Engineering) and two non-Science disciplines (Applied Linguistics and Business Studies), and so it will be possible to compare the results of this study with these. Again, however, he focuses on 4-word lexical bundles, and so it is only possible to compare his results with 4-word fixed collocation patterns.

Biber et al’s 12 structural categories for 4-word lexical bundles in academic prose (1999: 1014-1024) are shown in the first column on the left of Table 7.8 below. The second column explains the abbreviation of the form which is used in Figure 7.1 below. Hyland’s slightly simplified set of 7 categories (2008: 9), with more categories collapsed into “other”, is shown in the third column.

Lexical bundle structural category Biber et al. (1999)	**Biber et al. (1999)**	**Hyland (2008)**	**Example**
noun phrase + of-fragment	NP+OF	NP+OF	*the nature of the* *the size of the*
noun phrase + other post-modifier fragment	NP+OM	NP+OM	*the extent to which* *the united states and*
prepositional phrase + embedded of-clause fragment	PP+OF	PP+OF	*on the basis of* *in the case of*
other prepositional phrase fragment	OPP	OPP	*in this paper we* *at the same time*
anticipatory it + verb phrase/adjective phrase	IT+AP/VP	IT+AP/VP	*it is possible to* *it should be noted*
passive verb + prepositional phrase fragment	PV+PP	PV+PP	*shown in fig 6* *was added to the*
copula be + noun phrase/adjective phrase	BE+NP/AP	OTHER	*is the number of* *is likely to be*
(verb phrase +) that clause fragment	(VP+) THAT	OTHER	*can be seen that* *that it is not*
(verb/adjective +) to-clause fragment	(V/A+) TO	OTHER	*has been shown to*
adverbial clause fragment	ADV	OTHER	*as described in materials* *as a function of*
pronoun/noun phrase + be (+ ...)	PRON	OTHER	*there is no reason* *there has been a*
other expressions	OTHER	OTHER	*if and only if* *as well as the*

Table 7.8: Structural categories of lexical bundles Biber et al. 1999 and Hyland (2008)

In four structural categories there are noticeable similarities in the percentages of fixed collocational patterns in the present study and the percentages of the lexical bundles in Hyland's study (2008). The percentage of fixed collocational patterns with the form *prepositional phrase + embedded of-clause fragment* is higher in non-Science than in Science disciplines. This difference can also be seen in the differences between percentages of lexical bundles with this form in different disciplines in Hyland.

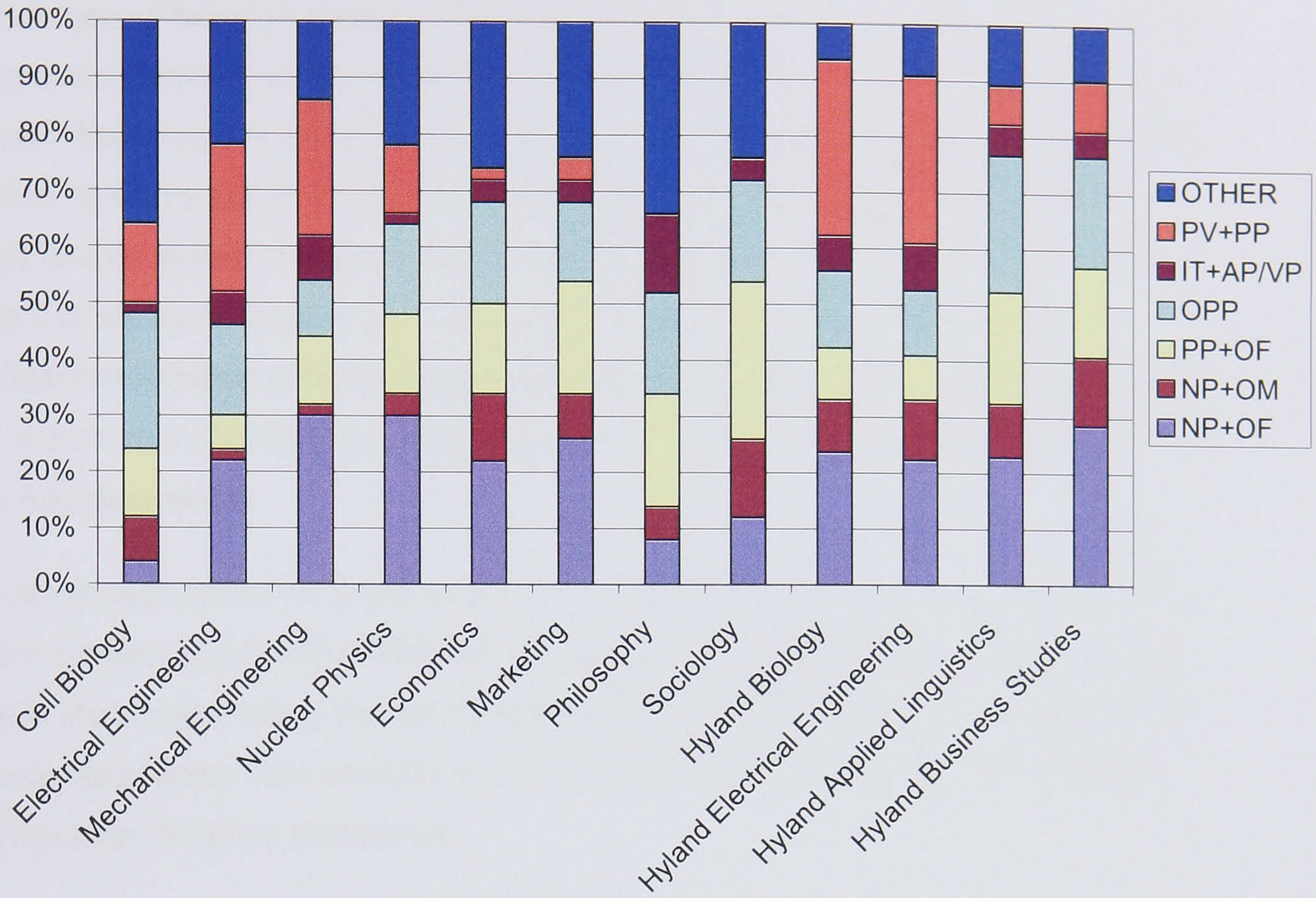


Figure 7.1: Comparison of the percentage of the top 50 4-word fixed collocational patterns in the present study with a particular form with the percentage of lexical bundles in Hyland (2008) with the same form

This broad agreement between the present study and Hyland's study can also be seen in the percentages of patterns with the form *other prepositional phrase fragment*. Again these are broadly higher in non-Science disciplines than in Science disciplines, with the one noticeable exception in this study being Cell Biology. The percentage of the form *anticipatory it + verb phrase/adjective phrase* is similar in all four of Hyland's disciplines, and is largely the same in the disciplines in the present study. One noticeable difference is in Philosophy, however, where the percentage of this form is markedly higher.

The fourth structure, that of *passive verb + prepositional phrase fragment* is, both in the present study and Hyland's study, much more common in Science than non-Science disciplines. In the present study it does not occur at all in Philosophy and Social Science.

The main differences between the percentages of forms in this study and the percentages of forms in Hyland's study are connected with noun phrases. There is a noticeable difference in the percentage of patterns with the form *noun phrase + of-fragment* between the disciplines in this study and the percentage of lexical bundles found in Hyland's study which have the same form. While the percentages are largely similar across the four disciplines in Hyland's study, the percentages in the present study do not reveal any observable tendency; in particular, a simple Science/non-Science difference cannot be seen. The percentage of patterns of this form occurring in Cell Biology, Philosophy and Sociology is less than for those in the other five disciplines.

The percentage of lexical bundles with the form *noun phrase + other post-modifier fragment* again occur with similar percentages in each discipline in Hyland's study, while in the present study the percentages are generally larger in non-Science than Science disciplines. One exception is Cell Biology which has a higher percentage than the other Science disciplines.

The final difference between the two studies is the percentage of forms which are categorised as *Other*. In Hyland's study the percentages are very similar in each discipline, while in the present study there is wide variation across disciplines. This is a consequence of the larger amounts of data in this study. Patterns which are less important in Hyland's study may occur more frequently in this study as a result of there being more opportunities for them to be used in the texts in the corpus and thus occur in the data.

7.3 Summary

This chapter has presented the main quantitative results of this study, relating to the first two questions. It first presented the relative quantities of fixed collocational pattern types and tokens occurring in the isolexical and isotextual versions. It then presented the distributions of the most frequent 3- and 4-word fixed collocational patterns in the isolexical and isotextual versions of the corpus, and showed that there is a remarkable stability in the rankings of these two sets of patterns. What is frequent isolexically has in this study been shown to be frequent isotextually.

This chapter then compared the results for the present study with those of Biber and Hyland, and showed that there are a limited number of isolexical fixed collocational patterns which are universally frequent in academic English regardless of genre. Some of Biber's isolexically fixed collocational patterns, i.e. the lexical bundles in his study, which are frequent in a genre corpus of textbooks in different academic disciplines, are also frequent in the present study, as are some of Hyland's isolexically fixed collocational patterns, i.e. the lexical bundles in his study, which are frequent in a corpus of academic English register.

The study will now move on to the exploratory study of the discourse function of the pattern *in the case of*. This is the most suitable pattern for detailed investigation in the remainder of this thesis as it is the only one to occur in all four data sets mentioned above. As was shown in Table 7.4, in the present study it occurs more than 100 times in all disciplines in both isolexical and isotextual versions of the corpus; it also occurs in every discipline in the textbooks in Biber (2006); it also occurs in every discipline in Hyland (2008), and, as shown in Table 6.9, it is also one of the three most significant 4-word academic written formulas in Ellis and Simpson (2005). The case study will apply the heuristic framework proposed in section 2.3.2 to this pattern in an attempt to reveal more about the relationship between its textual environment and discourse function.

Chapter 8 Exploratory case study: *in the case of*

8.1 Introduction

This section investigates how further studies of the functions of fixed collocational patterns in comparative studies across isotextual subcorpora might proceed. It examines the textual environments of *in the case of* using the heuristic framework of the four levels of collocation described in section 2.3.1. *In the case of* may not intuitively be the most interesting pattern to investigate, but it is the one suggested by the inductive, corpus-driven methodological approach developed in this thesis as the pattern most worthy of study.

The figures presented in this section are for occurrences in the isotextual corpus, allowing the use of this pattern across a similar number of communicative events to be compared. The low rank differences between frequent fixed collocational patterns found in the previous chapter mean that the two potential imbalances created by comparisons between data from texts of uneven length can be discounted. Because these results are from isotextual subcorpora, they will not be skewed by any discipline whose subcorpus contains many more texts than other disciplines. They will also not be overly influenced in favour of any discipline whose subcorpus, due to the shorter texts of which it is comprised, contains far fewer tokens than any other disciplines.

Examination of isotextual occurrences should be able to reveal whether there is a relationship between the uses of this pattern in research articles in one discipline and its uses in research articles in other disciplines. In Biber's lexical bundle framework *in the case of* is a "referential bundle with intangible framing attributes", while in Hyland's lexical bundle framework it is a "text-oriented framing signal". Both these studies assign these functions to *in the case of* and assume that this bundle functions in the same way regardless of the discipline in which it is used. This case study takes the opportunity to explore whether there might be other aspects of the observed use of *in the case of*, and which are common in and across different disciplines. Finally, since the functional frameworks of Biber and Hyland were

produced using register corpora, this case study should reveal more about the functions of this pattern in a corpus balanced for genre, i.e. the isotextual corpus of research articles in this study.

8.2 The most frequent collocation of *in the case of*

In the isotextual version of the corpus, *in the case of* occurs 3,661 times. To fully investigate this number of occurrences would be beyond the scope of this study, which has already involved the construction and analysis of two versions of a large corpus. This pattern does have a fairly restricted textual environment, however, and so an investigation can be made of this pattern with its next most frequent collocate. As shown in Figure 8.1 below, the most frequent word to occur immediately on the right of *in the case of* is *the,* which occurs 777 times. The next most frequent word is *a,* which occurs 288 times, only 37% of the frequency with which *in the case of* occurs with *the*. In the next position to the right of *in the case of the* there is very little regularity: only *brand* occurs more than ten times to the right of *in the case of the*, and these occurrences all come from a single text in the discipline of Marketing, as illustrated by Example 5.4 above.

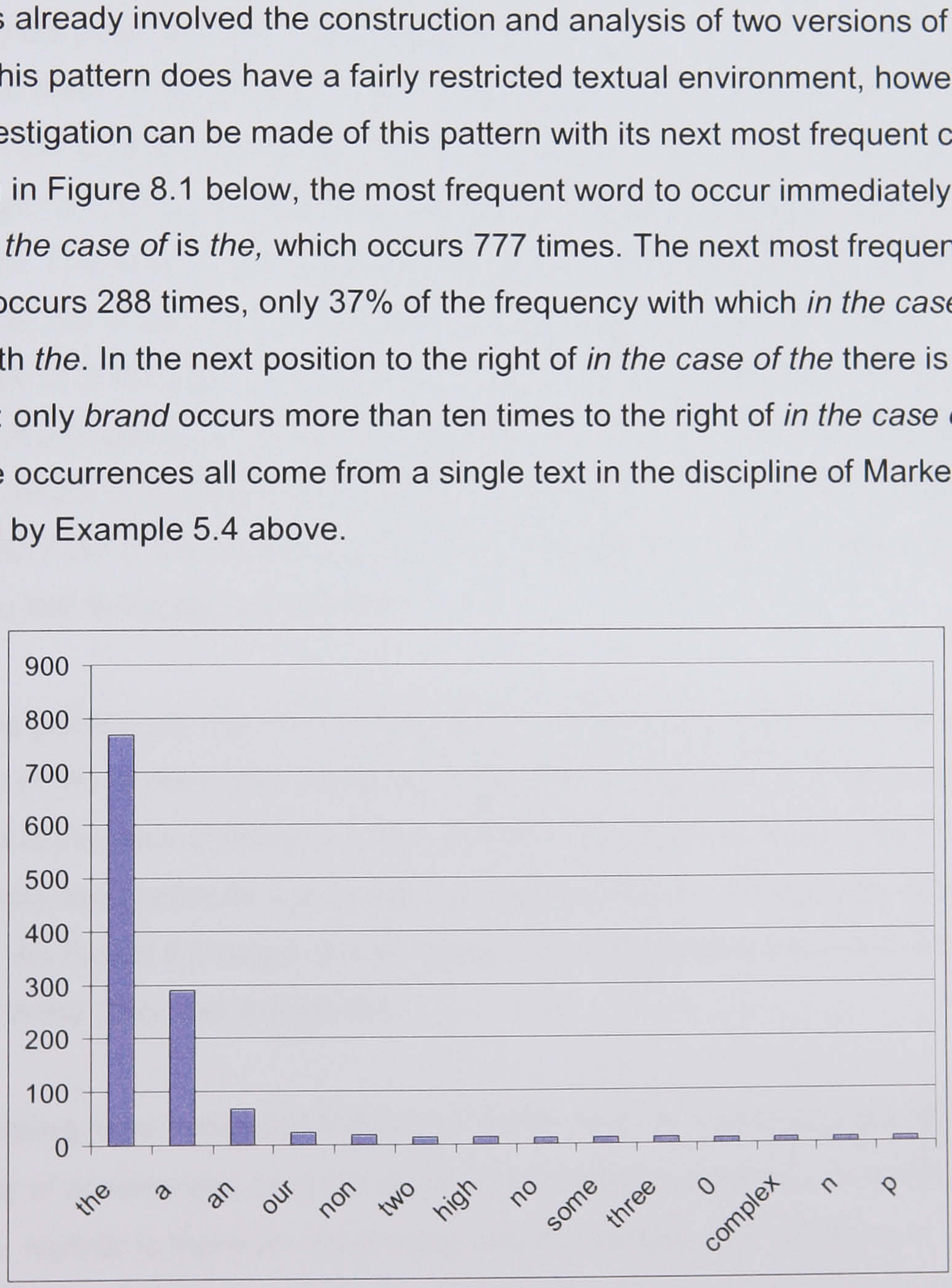


Figure 8.1: Words to the right of *in the case of* in the isotextual version of the corpus

This case study of *in the case of* with its most frequent collocate *the* will investigate its environments according to the levels of collocation suggested by both the Eastern European and empirical phraseological traditions described in section 2.3.1, namely those of lexis, syntax (namely position in sentence), semantics and pragmatics. Since this is a study of the genre of the research article, it will also investigate the additional element of textual location by adopting Biber et al's tool of the discourse map (Biber et al. 1998: 122).

Also, since this is a study of the functions of isotextual *in the case of the*, it will investigate the possibility of a relationship between the pragmatics of *in the case of the* with the other levels of collocation. Since these occurrences are found in a similar number of texts across disciplines, any relationship between the pragmatics of the pattern and these other levels of collocation should be revealed by the comparison. This should give a clearer indication of whether there is a relationship between the use of a pattern and its accompanying lexis. For example, do occurrences of *in the case of the* which are used to present findings from previous work exhibit a tendency to frame noun groups which contain lexis from semantic sets to do with results or findings? Or is there, for example, any relationship between the use of *in the case of the* and its position in the sentence, or between the use of *in the case of the* and its location in the text?

It should be pointed out here that this section examines only the most frequent collocation of the 4-word fixed collocational pattern *in the case of*. It does not investigate *in the case of the* as a fixed collocational pattern in its own right as it is only a 5-word fixed collocational pattern in Nuclear Physics, Philosophy, and Sociology. As Figure 8.2 below shows, *in the case of the* does not occur more than 100 times in the other five disciplines.

The first finding to be noticed in Figure 8.2 is that there is no observable difference in the number of occurrences of *in the case of the* between Science and non-Science disciplines. Neither is there an observable relation between the number of occurrences of *in the case of the* in a particular discipline and the VSD for that discipline. Nuclear Physics overall has the most occurrences even though its VSD is 0.54, (meaning that the isotextual subcorpus contains fewer tokens than those of

most other disciplines) and Cell Biology, which has a higher VSD of 0.8, has the fewest occurrences.

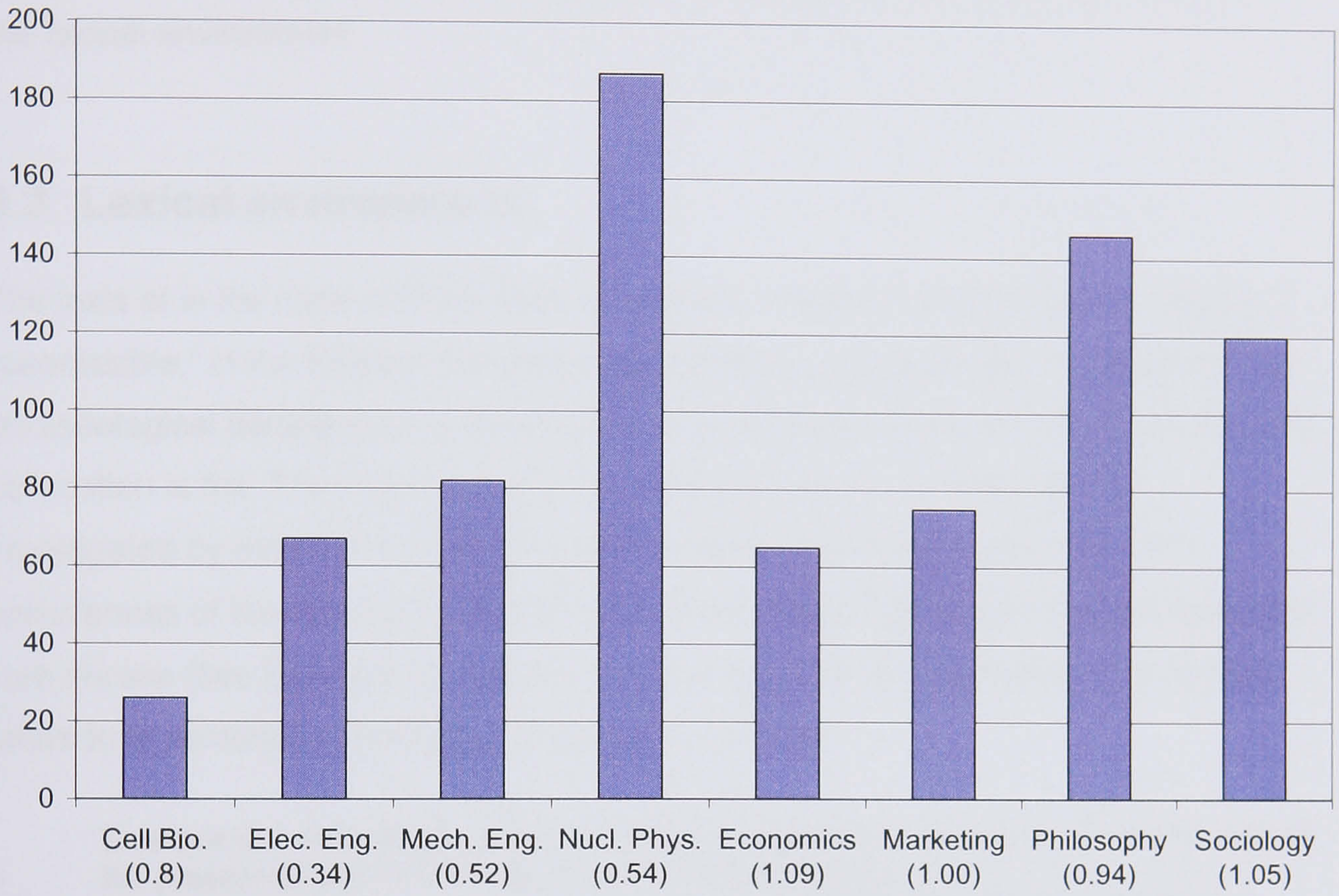


Figure 8.2: Occurrences of *in the case of the* in isotextual subcorpora

Two things can be said relating to the genre of the research article in different disciplines. The finding that *in the case of the* occurs most often isotextually in Nuclear Physics regardless of the number of tokens suggests that Nuclear Physics articles communicate more aspects of their research using this sequence than articles in other disciplines. This finding only concerns how texts communicate: it cannot be used to suggest any wider motivation of individual writers in Nuclear Physics articles since, as discussed above, Nuclear Physics articles are extensively co-authored.

The finding that the sequence occurs least of all isotextually in Cell Biology suggests that Cell Biology articles in this discipline communicate fewer aspects of their research using this sequence than articles in other disciplines. This same observation could have been made using the isolexical corpus: the sequence occurs

338 times in Nuclear Physics and 36 times in Cell Biology. However, the isotextual finding is needed in order to confirm that the isolexical quantities have not been distorted by the imbalance in the number of texts in the isolexical corpus. The next sections investigate individual occurrences of this sequence further, beginning with the lexical environment.

8.3 Lexical environment

The lexis of *in the case of the* is itself a constant. There is no other lexical variation "permissible," in the Eastern European perspective, or "observable," in the empirical phraseological perspective. *In the case of* is by definition fixed, and its most frequent collocation is *the*. The wider lexical environment *in the case of the* can be investigated by examining what it is which is framed by *in the case of the*. All occurrences of the string are followed by a noun group. There is one occurrence of a verb phrase (see Example 8.1 below), but this is in fact nominalised and in context refers to 'a sentence containing the words "it's raining."'

Example 8.1: Knowledge of the observational sort we are talking about in the case of the present-tensed 'It's raining' must be causally based.

(Philosophy 0080070480203)

An indication of the lexical environment of *in the case of the* is given by the head noun of the noun group framed by this sequence, which usually occurs between one and four words to the right. In Example 8.2 the head noun is *protein*, while in Example 8.3 the head noun is *control*. Some head nouns occur further to the right in noun groups with more premodification, such as in Example 8.4, but these are relatively rare.

Example 8.2: filament-forming ParM protein of plasmid R1

(Cell Biology 0070050221938)

Example 8.3: thermostat oven control

(Sociology 0100020480406)

Example 8.4: coupled-line phase-shifter topology

(Electrical Engineering 0030040520230)

This tendency for head nouns to occur within a limited number of words to the right of *in the case of the* is reminiscent of the collocational span around a node word described in section 2.3.1. This might suggest that different types of collocations collocate with each other, and that there are groups of lexical items which occur within a certain span of a fixed collocational pattern. The possibility of there being a separate collocational span for fixed collocational patterns is a question for further research and will be discussed in section 10.4 below. The lexical items to the right of *in the case of the* will, however, offer insights into the contexts of situation of the pattern as it is used in the different disciplines.

The relatively short length observed in the noun groups to the right of *in the case of the* meant it was possible to investigate its lexical environment using concordance lines. Even ambiguous noun groups, such as in Example 8.5 where the head noun could be *resonator* or *probes,* were short enough to be disambiguated using the concordance line. In this example, *probes* is not in fact part of the noun phrase introduced by *in the case of the*, but functions instead as the subject of the finite passive verb form *were placed*:

> Example 8.5: In the case of the dual-mode resonator probes were placed in the sidewalls for the WGE mode (Electrical Engineering 0030020400127)

All head nouns occurring with *in the case of the* are shown in Appendix 20. A summary of these is shown below in Table 8.1 which shows the extent of the lexical environment framed by the string in each discipline, and also gives an indication of how much that environment is shared with other disciplines. The table first shows how many nouns are framed by the string, and then how many of these nouns are also framed by the string in other disciplines. This latter figure is given as a percentage figure in the third row. The next rows show the actual nouns, together with their distribution across disciplines. So, as an example, *in the case of the* frames 93 different nouns in Sociology, 15 of which it shares with other disciplines, so that 16.1% of the lexical environment which this string frames in Sociology is also framed by it in other disciplines.

	Cell Bio. (0.8)	Elec. Eng. (0.34)	Mech. Eng. (0.52)	Nucl. Phys. (0.54)	Econ-omics (1.09)	Market -ing (1.00)	Philos -ophy (0.94)	Socio-logy (1.05)	Occurr-ences of each noun	Number of disciplines
Total nouns	26	42	56	102	44	43	104	93		
Total shared nouns	2	8	16	18	15	9	15	15		
% of total nouns	7.7	19	28.6	17.6	34.1	21	14.4	16.1		
agents					1		1		2	2
analysis			1	1					2	2
categories						3	4		7	2
components		1			1				2	2
concepts							1	1	2	2
control		1						1	2	2
data			3	1				1	5	3
equation				1	1				2	2
example			1		1	3	2		7	4
experiment				1			1		2	2
film		4	1						5	2
former						1	1	3	5	3
formulation				1			1		2	2
function				3	2				5	2
functions			1	1					2	2
grid		3						1	4	2
industry			1		2	2		3	8	4
latter						2		2	4	2
material		1	1						2	2
method			1	1					2	2
model			5	13	6	5		1	21	5
models				1		1			2	2
number			1				1		2	2
particle			2	2					4	2
problem			1		2		2		5	3
programme					1			1	2	2
rate				1	1				2	2
reduction				1			1		2	2
results			1		1				2	2
right							1	1	2	2
set		1	2						3	2
solution				1			1		2	2
state	1					1			2	2
states				5	2			3	10	3
structure		2			1				3	2
studies						1		1	2	2
survey					1			1	2	2
system		3	1	4			1	1	10	5
theory				2			3		5	2
thesis							1	1	2	2
treatments	1				2				3	2
waves			2	1					3	2
Total shared occurrences:	2	16	25	41	25	19	22	22		

Table 8.1: Head nouns of noun groups framed by *in the case of the* in the isotextual version of the corpus

The discipline which shares the most of the lexical environment of this string, 34.1%, is Economics. The subcorpus for Economics also has the highest VSD (1.09) and so the reason it shares the most environment with other disciplines could be that its isotextual subcorpus contains more tokens than the other subcorpora, and thus it contains more lexis to share. However, the amount of shared lexical environment appears to be independent of subcorpus size when percentages for other disciplines are examined. The discipline with the lowest VSD, Electrical Engineering (0.34) shares 19% of its environment, while a discipline with a much higher VSD, Cell Biology (0.8) shares only 7.7% of its environment. The fact that most disciplines share no more than 20% of their lexical environment with other disciplines suggests that the lexical environment framed by *in the case of the* depends on the communicative purposes of the text in the discipline in which it is used. This shows that the lexical environment at least is largely discipline-specific.

8.4 Semantic environment

Table 8.1 suggests there is little regularity in the semantic environment of this sequence, and so this level will not be investigated further in this thesis.

8.5 Syntactic position

In the case of the can be used in three positions in a sentence: initial, near-initial, and non-initial. Initial position is at the start of the sentence, as in Example 8.6 and Example 8.7 below:

> Example 8.6: In the case of the low-lying 2+ states, the disagreement is almost certainly related to the existence of important effects which go beyond the mean field theory adopted here.
>
> (Nuclear Physics 0090047260101)

> Example 8.7: In the case of the concepts of symbolization and generalization, thirdly, the former refers to the exchange-value of money as well as the symbolic utility of language
>
> (Sociology 0100040530306)

Near-initial position is defined as before the main verb, i.e. after an initial connective or as part of the subject of the main clause, as in Example 8.8 and Example 8.9 below:

> Example 8.8: However, at least in the case of the regular synaptotagmin 7 variant, we only obtained evidence for a regulatory effect in recycling upon stimulation of exocytosis by strong, unphysiological K+ depolarizations.
>
> (Cell Biology 0070050222003)

> Example 8.9: Knowledge of the observational sort we are talking about in the case of the present-tensed 'It's raining' must be causally based.
>
> (Philosophy 0080070480203)

Non-initial position is anywhere after the main clause, as in Example 8.10 and Example 8.11 below:

> Example 8.10: The dependence of Young's modulus on is slightly more complex than in the case of the aggregate modulus, as shown in Eq. (32).
>
> (Mechanical Engineering 0060050370316)

> Example 8.11: The heterogeneity parameters in Table 2 indicate that there is some heterogeneity in the intrinsic brand preferences in the case of the most general model although the variances are not very large in magnitude.
>
> (Marketing 0050050200410)

The relative proportions of occurrences of *in the case of the* in different positions are shown in Figure 8.3 below. There are few noticeable patterns in the syntactic distribution of occurrences of *in the case of the*. Occurrences in initial position show little relationship to Science or non-Science disciplines, nor to the length of the texts in the subcorpora for that discipline. Less than a third of occurrences are in initial position in Nuclear Physics, while over half are in initial position in Mechanical Engineering, even though the subcorpora for these disciplines have similar VSDs (0.52 and 0.54). Economics and Sociology, on the other hand, whose subcorpora also have similar VSDs (1.09 and 1.05) show virtually no difference in the percentage of sentence-initial occurrences. These two observations suggest that the difference in the proportion of sentence-initial uses of *in the case of the* is not likely to be a result of differences in text lengths in these disciplines.

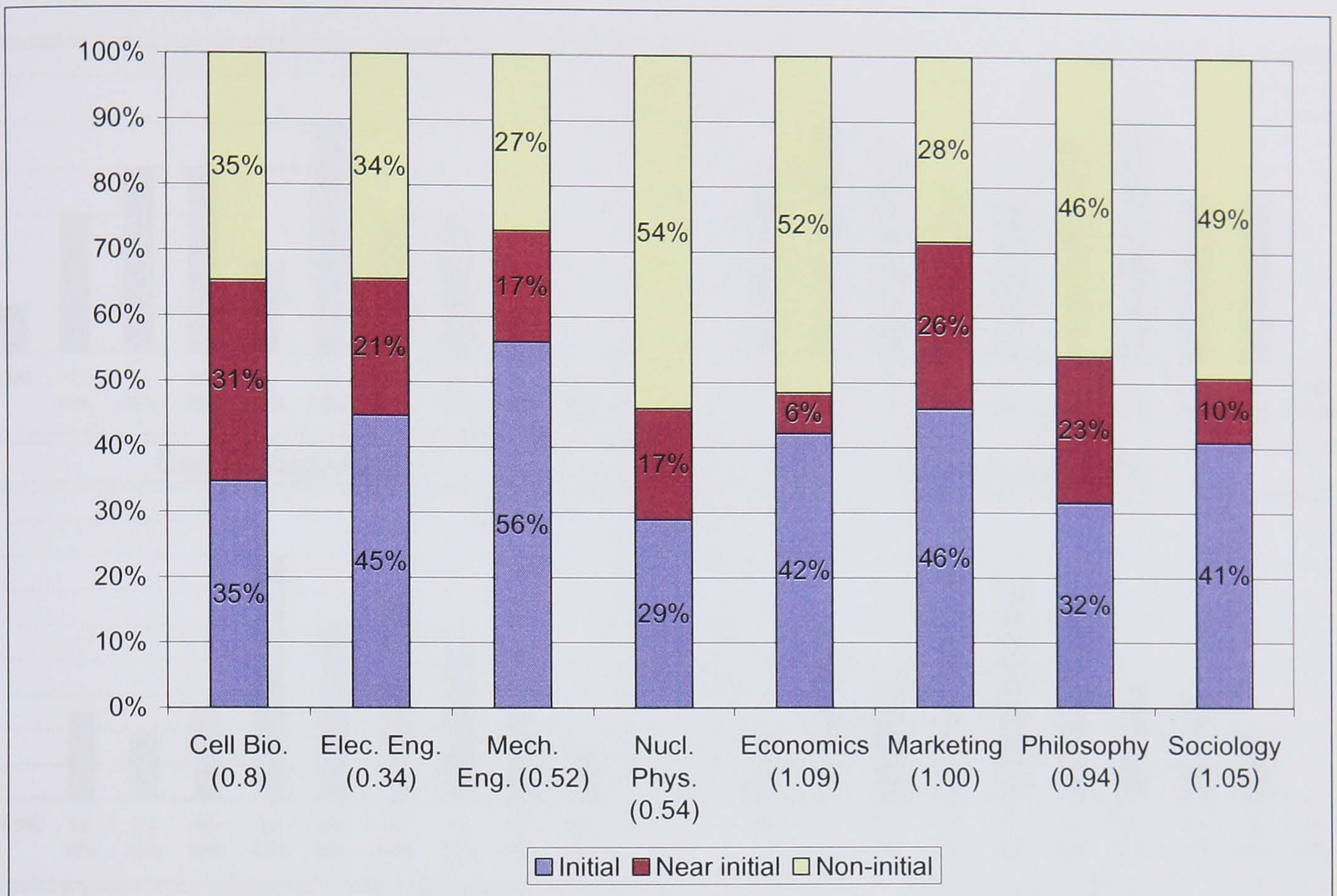


Figure 8.3: Sentence position of *in the case of the* in each discipline in the isotextual version of the corpus

There is also a lack of any relationship noticeable in the percentage of occurrences which are non-initial and near-initial, other than the larger the percentage of occurrences which are sentence-initial, the smaller the percentage of occurrences which are initial. The one exception to this is Cell Biology, in which occurrences are evenly split between each of the three positions.

8.6 Textual location

The textual location of *in the case of the* can be shown in two ways. First, it can be seen in percentage terms: the further from the beginning of the text it occurs, the higher the percentage figure. This gives an objective view of the physical location of the string in the text rather than its functional location in the discourse. The percentage locations of each occurrence of *in the case of the* in each subcorpus are shown in Figure 8.4 below. The graph for each discipline is divided into ten deciles, and each bar shows the number of occurrences of *in the case of the* in each decile.

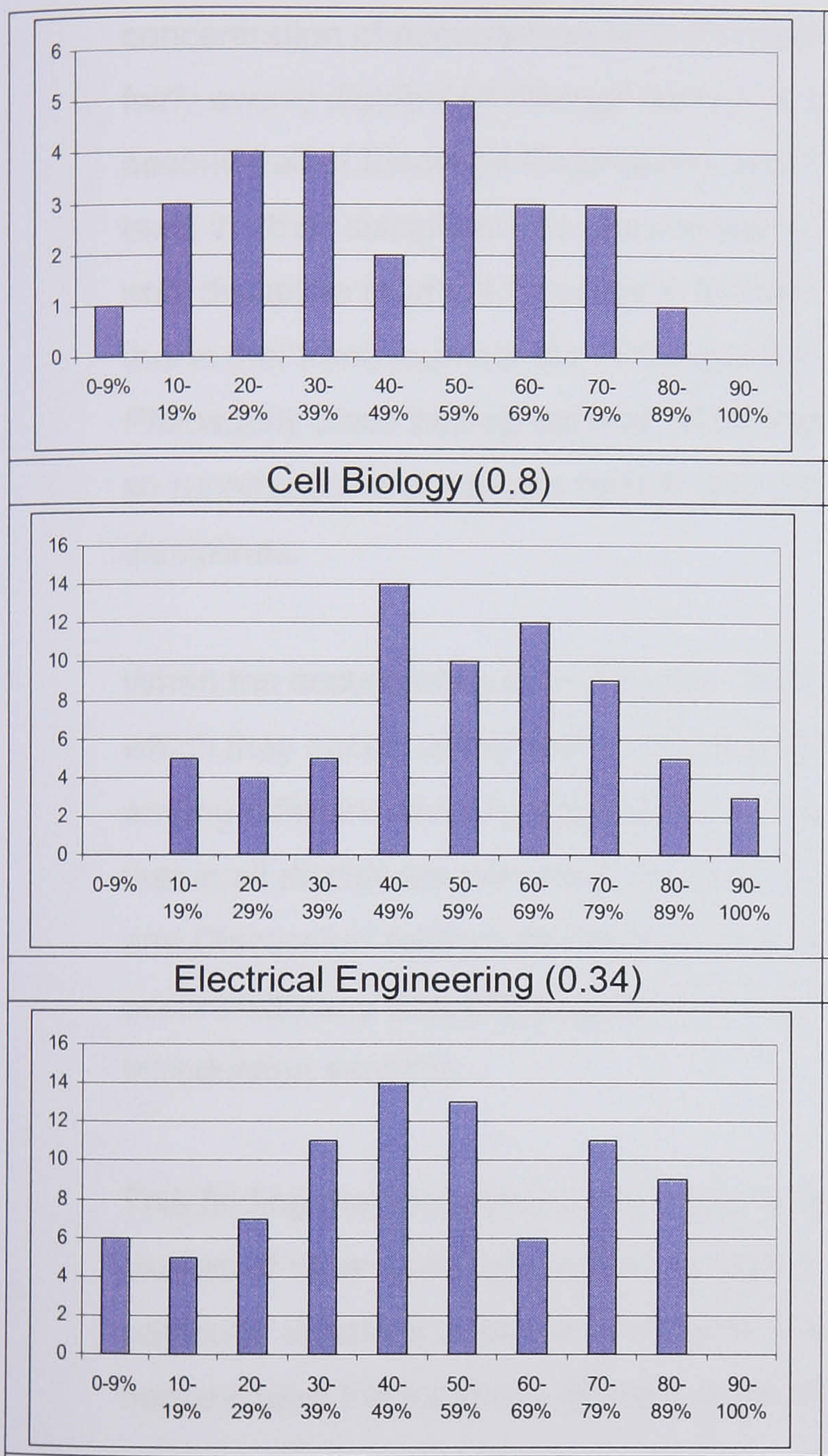


Figure 8.4: Textual location of *in the case of the* in each isotextual subcorpus, by percentage of text

It can be seen in Figure 8.4 that on the percentage count there is little indication of a concentration of occurrences of *in the case of the* in any location. The sequence is fairly evenly distributed through texts in Economics, tends to occur more in the second half of Electrical Engineering and Philosophy, and irregularly throughout texts in other disciplines. Its occurrence in Philosophy is unusual since this is the only discipline in which it occurs in the last 90% of a text. The likely explanation for this is that three journals the *Philosophical Review*, the *Philosophical Quarterly*, and *Philosophy* place bibliographical references in footnotes rather than endnotes, and so running prose continues right to the end of the text, unlike articles in other disciplines.

When the occurrences of *in the case of the* are examined in relation to the section in which they occur, rather than in terms of the percentage decile, their distribution among different sections appears more regular. It can be seen in Figure 8.5 below that in all disciplines the string occurs most often in the Body section or the Results and Discussion section. Nuclear Physics is the only discipline in which the string occurs with any sizeable frequency in any other section, as it occurs 20 times in Introduction sections.

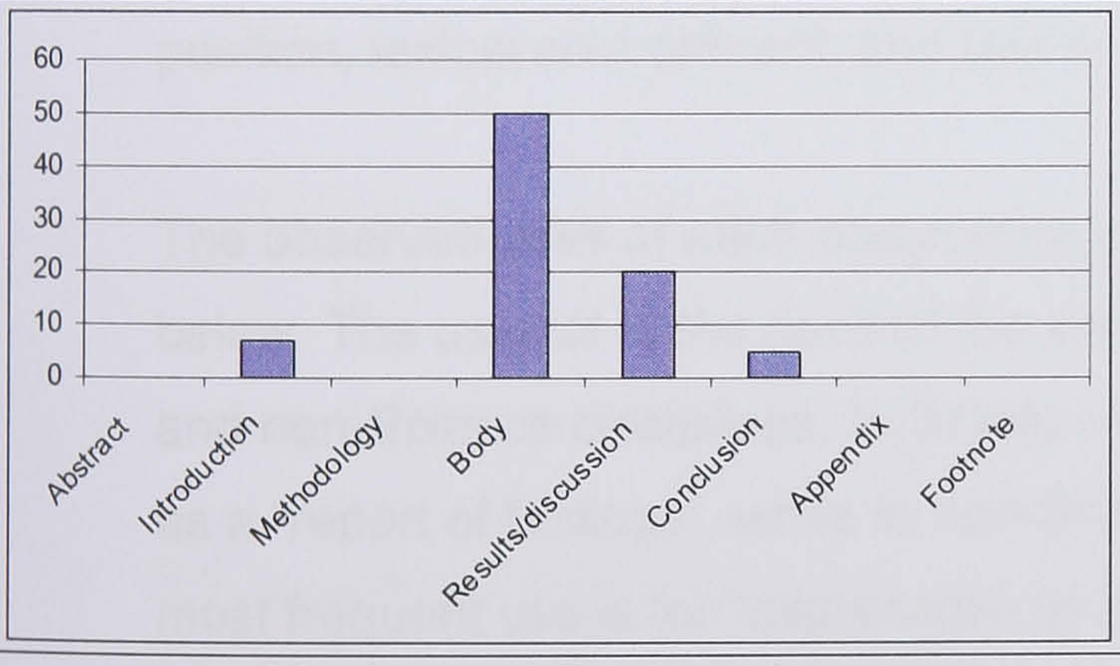
This finding may not seem particularly revealing with *in the case of the*, but in future studies of other fixed collocational patterns using this methodology, concentrations in particular locations which had not been expected might be revealed. Figure 8.5 below shows the locations of *in the case of the* in each text section:

Cell Biology

Economics

Electrical Engineering

Marketing



Mechanical Engineering

Philosophy

Nuclear Physics

Sociology

Figure 8.5: Textual location of *in the case of the* in each isotextual subcorpus, by section

Having discussed the lexical, semantic and syntactic environments of this string, as well as its textual location, we now move on to an investigation of its pragmatic discourse function.

8.7 Pragmatic function

In order to examine the more detailed pragmatic use of the sequence, all occurrences of *in the case of the* were examined with reference to their original context. This was done according to the procedure described in section 5.5.4 rather than relying on concordance lines. Uses of this pattern appear to fall into four main categories: description of procedure, findings from previous work, report of findings, and explanation of topic. The uses of *in the case of the* within each category will be described briefly, and representative examples given from each discipline. Some examples contain more than one sentence in order to make the context of the use of the sequence clearer and show that it belongs in the category assigned to it. The final section in this case study will then investigate the relationship between the observed use of *in the case of the* and the previous criteria, namely its syntactic position, lexical environment, and text location.

The observed uses of each occurrence in each discipline are shown in Figure 8.6 below. The uses of *in the case of the* seem to show some variation between Science and non-Science disciplines. In Science disciplines the sequence is most often used as a "report of findings", while in non-Science disciplines, apart from Marketing, its most frequent use is for "explanation of topic". This might be the result of the influence of the number of tokens in each isotextual subcorpus. Non-Science disciplines have higher VSDs and so contain more tokens, and so their longer texts might contain more explanation of the topics which they are about.

The variation between Science and non-Science disciplines might on the other hand reflect epistemological practices between Science and non-Science disciplines. It may be that since knowledge in Science papers is created more by experimentation than in non-Science papers, then each Science text will contain more reports of findings. However, there are some similarities between non-Science and Science

disciplines. In Marketing and Electrical Engineering *in the case of the* is used for reporting of findings a similar number of times in the same number of texts.

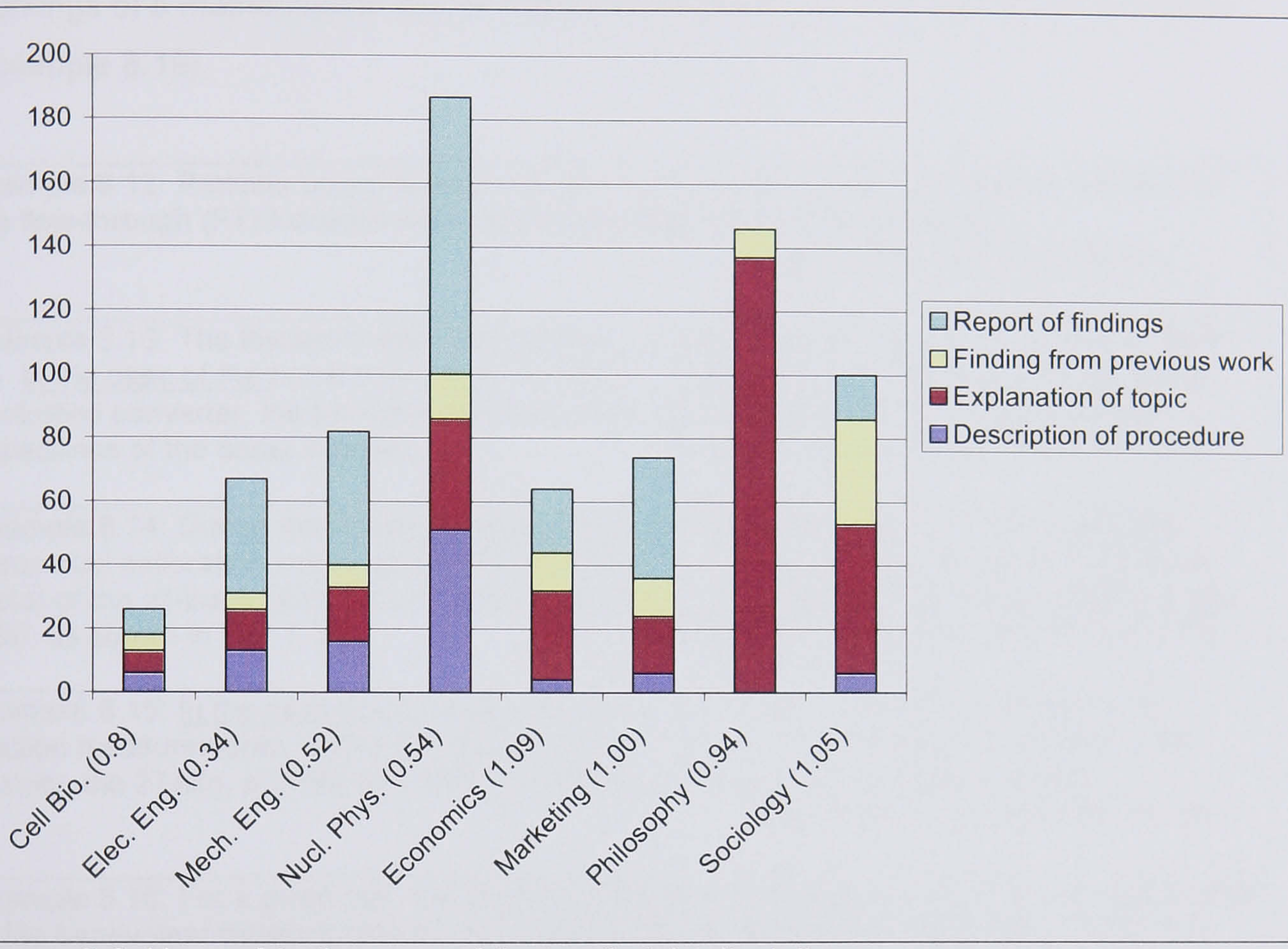


Figure 8.6: Uses of *in the case of the* in all eight disciplines

The next sections present examples of each function from each discipline.

8.7.1 Description of procedure

This use of *in the case of the* frames an aspect of the process by which the findings presented in the article were obtained. As shown by the examples in Table 8.2 below, the sequence is used to frame a range of methodological approaches. In Science subjects it is mostly used to frame an aspect of an experimental procedure, as in Cell Biology (Example 8.12), Mechanical Engineering (Example 8.14), and Nuclear Physics (Example 8.15). In Electrical Engineering (Example 8.13), however, it more often is used in descriptions of how a device works or is tested. In non-Science subjects the pattern is used when describing the design of an experiment, such as reasons why a sample population is chosen (in Sociology Example 8.18), the instructions given to game participants (in Economics Example 8.17), and

questionnaire design (in Marketing Example 8.16). No occurrences of this use could be identified in Philosophy. A small number of occurrences in some of the more mathematical journals in Economics *in the case of the* is used when describing the workings of a mathematical model expounded, developed, and applied in the article (Example 8.19).

Example 8.12: Proteins bound to each column were applied to the next column, except that the flow-through (FT) fractions were used in the case of the heparin column. (Cell Biology 0070011630601)
Example 8.13: The thyristors were also used to control the inrush current on converter start-up. In the case of the boost converter, the thyristor bridge was operated as a simple phase controlled converter, the turn-on angle being determined by the DC link voltage and the impedance of the boost inductor. (Electrical Engineering 0030061500606)
Example 8.14: During the forming process, tests were performed at a uniform strain rate; theoretical calibration curves were obtained from a finite element (FE) analysis. The friction factor of the oil-base lubricant was about 0.30; in the case of the water-base lubricant, it was 0.60, as shown in Fig. 1. 2.2. (Mechanical Engineering 0060032170710)
Example 8.15: In the case of the 60Ni(n, p)60mCo and 61Ni(n, x)60mCo reaction cross section measurements, where the irradiation time was too short to build up sufficient 24Na activity, the 27Al(n, p)27Mg and 56Fe(n, p)56Mn reactions were used as monitors. (Nuclear Physics 0090047300301)
Example 8.16: For a given pair, the average per-letter credit was always 50 cents. In the case of the benevolent dictators (and in their case alone), there was only one phase. They were informed of their role as allocator, of their $5 fixed fee in addition In two cases, however, certain information was not provided until a later phase. (Economics 0020080900422)
Example 8.17: To assist subjects with their understanding of the product category, sample products in the category were noted. For example, in the case of the home security products category, reference products included smoke detectors, carbon monoxide detectors, fire extinguishers, and home security alarms. (Marketing 0050030380304)
Example 8.18: In the case of the 'greens' this did not involve much selection though I attempted to ensure that all the main types of campaigning organisation were represented. This was more difficult in the case of the feminists for the reasons which I mentioned earlier, though I was able to interview women representing diverse strands of feminism. (Sociology 0100020470203)
Example 8.19: For example, in the case of the CES production function these elasticities are ax[x/F(x)]Iota, where ax represents the various distributional parameters and Iota is the substitution parameter. (Economics 0020080900422)
Table 8.2: Examples of *in the case of the* used for "description of procedure"

8.7.2 Finding from previous work

This use of *in the case of the* is the fairly straightforward framing of a previous research finding that is introduced into the text.

Example 8.20: In the case of SFCs a few protein domains have been described: the RS-domain and the RNA-recognition motif for SR proteins and CrkRS (Caceres et al., 1997; Gama-Carvalho et al., 1997; Ko et al., 2001); the `forkhead-associated' domain in the case of the protein phosphatase-1 regulator NIPP1 (Jagiello et al., 2000); (Cell Biology 0070021161504)
Example 8.21: It is important to notice that vibration excitation Qvib measured by Moran[12] in the case of the O2+/O2 system is quite low and therefore does not really affect the ion swarm data. (Electrical Engineering 003030940110)
Example 8.22: MRI or CT based studies (e.g. Nordt et al., 1999; Eckstein et al., 2000) have also been used to evaluate the deformation of soft tissues or motion under a particular loading. However, the time required to acquire a full image limits its use to quasi-static applications or highly reproducible motion. Overall, even if those techniques have shown better results than external markers, they are still limited in terms of accuracy for the current application. Furthermore, their sample rate is generally limited to 30 frames per second at the most in the case of the fluoroscopy. (Mechanical Engineering 0060050370707)
Example 8.23: In the case of the 2SC phase, this is studied in Ref. [26]. It is shown that the three gluons of the unbroken SU(2)c, unlike the other five gluons of the original SU(3)c, do not experience the Debye screening in the far infrared region. (Nuclear Physics 0090047290209)
Example 8.24: Riezman [1985], Yi [1996], and Burbridge et al. [1997] show that in many cases, because of international redistributive effects of preferential arrangements, multilateral free trade may not evolve in equilibrium in frictionless coalition games (even in the presence of international side payments, in the case of the latter paper). (Economics 0020091170207)
Example 8.25: One exception is the research by Erdem (1998), who examined household purchase data collected alter the brand extensions had been introduced and demonstrated that quality perceptions transfer between umbrella-branded products in the case of the companion categories of toothpaste and toothbrushes. (Marketing 0050030380304)
Example 8.26: Malament then finally proves, as his main result, that in the case of the Kerr solution, there is no generalized criterion of non-rotation that satisfies both adequacy conditions. (Philosophy 0080010540101)
Example 8.27: In the case of the Mardu, of Australia, it is observed that whatever trade that occurs among them takes place as either ritual activities during big meetings, gift exchange between friendly kin when small bands meet, or 'as part of obligations owed to certain close kin and affines' (Tonkinson, 1991: 53). (Sociology 0100030500103)
Table 8.3: Examples of *in the case of the* used for "finding from previous work"

The pattern may frame a substance, phenomenon, technique, or construct about which more was learned in the earlier work and which is relevant to the current paper. The examples in Table 8.3 above illustrate this use of the sequence in each discipline.

8.7.3 Report of findings

This identified use of the sequence is to report the findings of the research in the current paper. Again, no occurrences of this use could be identified in Philosophy.

Example 8.28: Fluorescent labels of membranes (FM4-64), lipid (Nile Red, not shown) and the cytosol (targeted GFP; 10,000 kDa dextran-dye conjugates; Figs 1, 4) showed by dye accumulation, or in the case of the cytosolic markers, by dye exclusion, that the BIA consists of lipid as membranes. (Cell Biology 0070021161308)
Example 8.29: When 40 impulse voltages are supplied the search coil EMF values are higher than in the case of the 30 voltages. But the search coil EMF ratio values show the same magnitude as in the case of the 30 voltages. (Electrical Engineering 0030061510401)
Example 8.30: The following observations can be made: (a) the vortex trace becomes closer (vertical distance) to the hydrofoil as the gap size is increased; (b) the trajectories in the 0.6 and 1.4 mm gaps are continuous and do not show a bump, i.e., a change in slope of the vortex trace that is clearly evident in the case of the 2.6 mm gap (Fig. 11(c)); (Mechanical Engineering 0060011240420)
Example 8.31: For the five states with similar angular distributions the secondary maxima shift a little with increasing excitation energy (as indicated by the dashed line in Fig. 3), but lie within the region 29° to 41°. We note that the secondary maxima appear to be less strong in the case of the 14.1 and 14.6 MeV states. (Nuclear Physics 0090047280102)
Example 8.32: Finally, column (5) of Table 1 states the percentage of allocations that conform exactly to the theory, i.e., the percentage of observations that equal the point predictions in the case of the discretionary treatments or the interval predictions in the case of the standard and double treatments. (Economics 0020080900422)
Example 8.33: When addressability is *i perfect (a 1), we have that p r. In other words, *i when all consumers are addressable, the posted prices become irrelevant for consumer decision making and all consumers buy at the addressable prices. In the case of the choice of addressability, we find a similar result to that of Proposition 2 when addressability comes at zero marginal cost. (Marketing 0050050210205)
Example 8.34: Though there was indeed some association between smoking and the image of being an adult it was not a very strong link (not statistically significant in the case of the girls) with less than one in eight of the girls who were regular smokers agreeing with the statement that 'Smoking is part of being grown up' and about one in six of the boys who were regular smokers doing so. (Sociology 0100040520109)
Table 8.4: Examples of *in the case of the* used for "report of findings"

8.7.4 Explanation of topic

In this last category, the sequence is used to convey information as part of the general expository discourse of the text, as shown in Table 8.5 below.

Example 8.35: In the case of the melanosome, the melanophilin-dependent coupling of Rab27a to myosin-Va drives the biological end-point of this organelle, which is its myosin-Va-dependent accumulation at dendritic tips. (Cell Biology 0070030040401)
Example 8.36: The filled circles and the open squares represent the films with and without metallic Hf layer predeposition, respectively. In the case of the films with metallic Hf layer predepositon, the horizontal axis represents the reactive sputtering time after the metallic Hf metal layer predeposition. (Electrical Engineering 00300508501 18)
Example 8.37: Feedback is an aspect of the system structure and also of the organization. In the case of the man±machine interface, human reliability increases if the same information is received by at least two sensory organs. (Mechanical Engineering 0060032171112)
Example 8.38: These two recent developments have been responsible for a renewed focusing of attention on the subject of relativistic shock fronts. In each case, the observations imply that the photon density in the source is so large, that the gamma-rays would be subject to strong absorption due to photon–photon interactions, were not the source moving in our direction with a speed close to that of light. The implied Lorentz factors are about 50–100 in the case of the active galaxies, and several hundreds in the case of gamma-ray burst sources. (Nuclear Physics 0090010250801)
Example 8.39: Instead, a large part of the causation appears to run in the other direction, from growth to saving. This is most evident in the case of the East Asian economies, which had high growth rates long before they had exceptionally high saving rates. (Economics 0020080900302)
Example 8.40: Applied to the advertising context, the specific activation mechanism suggests that omitting an ad element should improve later recall specifically along dimensions related to that element. For example, in the case of the Olympus advertisement that exemplifies overt omission, the omitted element (i.e., a missing camera on the man's chest) is related only to the product category (camera). (Marketing 0050030390204)
Example 8.41: There is a regular, though not invariable, connection between the appearance of heavy clouds and the falling of rain; and there is the same kind of connection between the command 'Sit!' and the dog's sitting. It is true, however, that in the case of the dog the regularity has been produced by a process of conditioning—a process in which the desired regularity is itself created by a regular combining of behaviour with sanctions. (Philosophy 0080050780102)
Example 8.42: We prefer the term representation, however, because it indicates not only formal involvement, but involvement for the purposes of representing the interests of a typically excluded and voiceless constituency. Thus, in the case of the homeless, representation provides a level of institutional voice for the homeless population by enabling the SMO to have input in policy decisions that affect homeless people. (Sociology 0100040520109)
Table 8.5: Examples of *in the case of the* used for "explanation of topic"

Having given examples of the four identifiable pragmatic uses of this sequence, the case study continues by investigating possible relationships between its use with the other levels of collocation, namely lexical environment, syntactic location, and textual location.

8.8 Investigating the relationship between pragmatics, lexis, syntax, and text location

8.8.1 Pragmatics, lexis and text location

Table 8.6 below relates the lexical environment of *in the case of the* to the collocational levels of textual location and pragmatics.

	Explanation of topic	Finding from previous work	Report of findings	Description of procedure
Introduction	function (Nucl) states (Econ) states (Econ)	categories (Mktg) film (Mech) former (Soc) industry (Econ)		system (Nucl)
Body	categories (Phil) categories (Phil) former (Phil) former (Soc) function (Econ) industry (Econ) industry (Mktg) model (Econ) model (Econ) model (Mktg) model (Mktg) model (Soc) system (Phil)	categories (Phil) former (Mktg) industry (Mech) industry (Mktg) industry (Mktg) states (Soc) states (Soc) states (Soc) system (Elec)	categories (Mktg) data (Nucl) function (Nucl) industry (Mktg) model (Nucl) model (Nucl) model (Nucl) model (Mktg) model (Mktg) states (Nucl) system (Elec) system (Elec) system (Mech)	former (Soc) function (Econ) model (Mech) model (Mech) model (Mech) model (Mech) model (Nucl) model (Nucl) model (Nucl) model (Nucl) model (Nucl) model (Nucl) model (Nucl) model (Nucl) system (Nucl)
Results / Discussion	data (Mech) system (Soc)	model (Nucl)	categories (Mktg) data (Mech) data (Mech) data (Soc) film (Elec) film (Elec) film (Elec) film (Elec) function (Econ) model (Mech) model (Econ) model (Econ) model (Econ) model (Mktg) states (Nucl) system (Nucl) system (Nucl)	states (Nucl)
Conclusion		categories (Phil)	industry (Mktg) model (Econ) states (Nucl) states (Nucl)	
Appendix	model (Nucl)			

Table 8.6: The relations between pragmatics, lexis, and text location of *in the case of the* for the 10 most frequent nouns

Table 8.6 shows the 10 most frequent nouns to occur to the right of the sequence, the section in which they occur, and its use. It can be seen that that the majority of occurrences are either in the Body section or the Results and Discussion section. There are relatively few occurrences of frequent lexis in Introduction and Conclusion sections. The sequence is used for all four functions in the Body section, while in the Results and Discussion section the majority of uses are for "report of findings." In the Body section, the most frequent lexis is *model,* which is particularly commonly used for "explanation of topic", "report of findings", and "description of procedure", although this last use is almost exclusively in Nuclear Physics.

The lexical environments of the sequence when used in the Body section give some idea to the context of situation in this section with each function in the different disciplines:

Explanation of topic:	categories, former, function, industry, model, system
Finding from previous work:	categories, former, industry, states, system
Report of findings:	categories, data, function, industry, model, states, system
Description of procedure:	former, function, model, system

There is a considerable amount of overlap between these lexical items and those occurring with this sequence with the "report of findings" function in the Results section:

Report of findings:	categories, data, film, function, model, states, system

8.8.2 Pragmatics and syntactic location

The next comparison to be made between the different collocational levels is between the discourse function and the syntactic location, namely the position in the sentence. The occurrences in terms of these two levels in each discipline are shown in Figure 8.7 below. There appears to be little relationship between the location of the string and its discourse function in any of the disciplines, however.

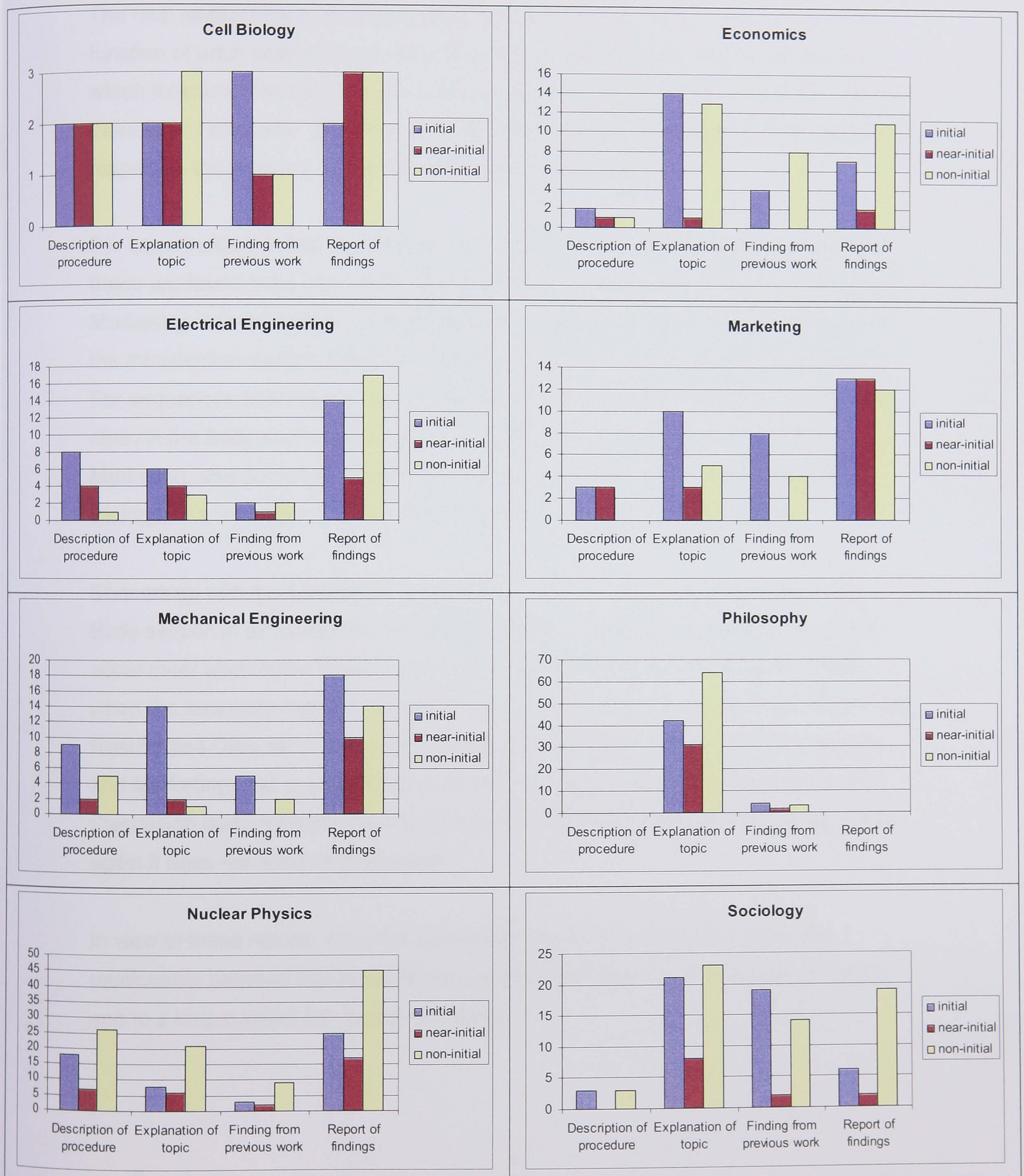


Figure 8.7:The relation between sentence location and use of *in the case of the* in each isotextual subcorpus

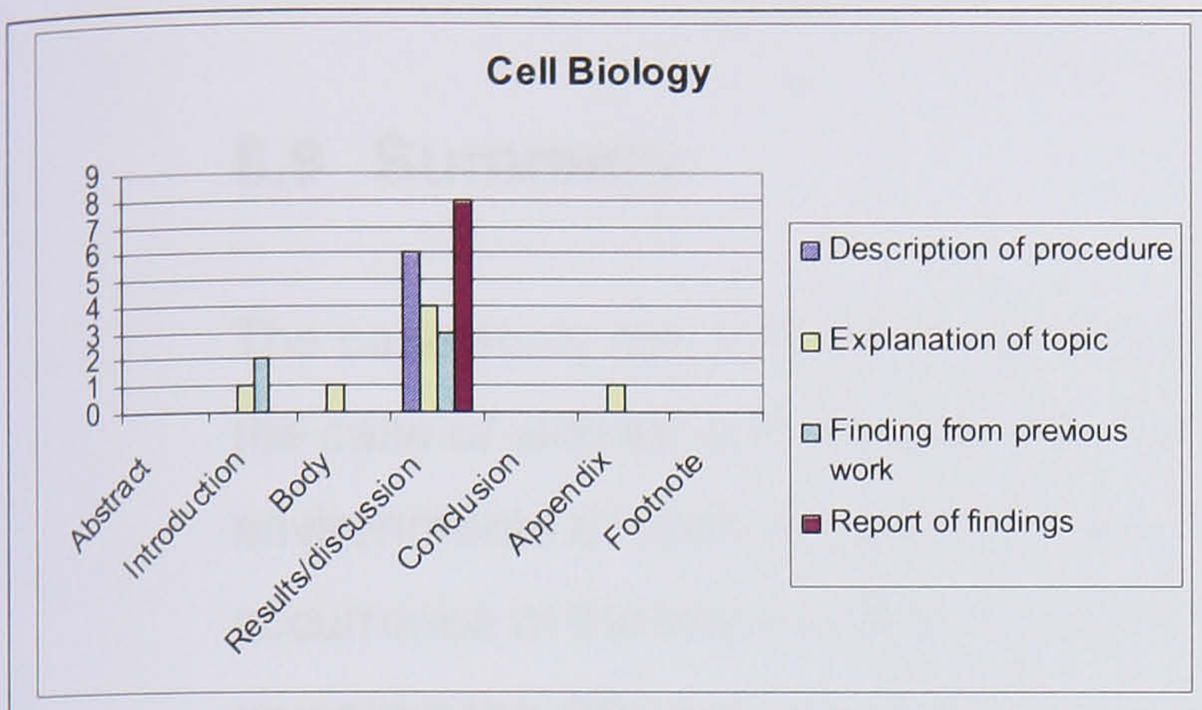
8.8.3 Pragmatics and textual location

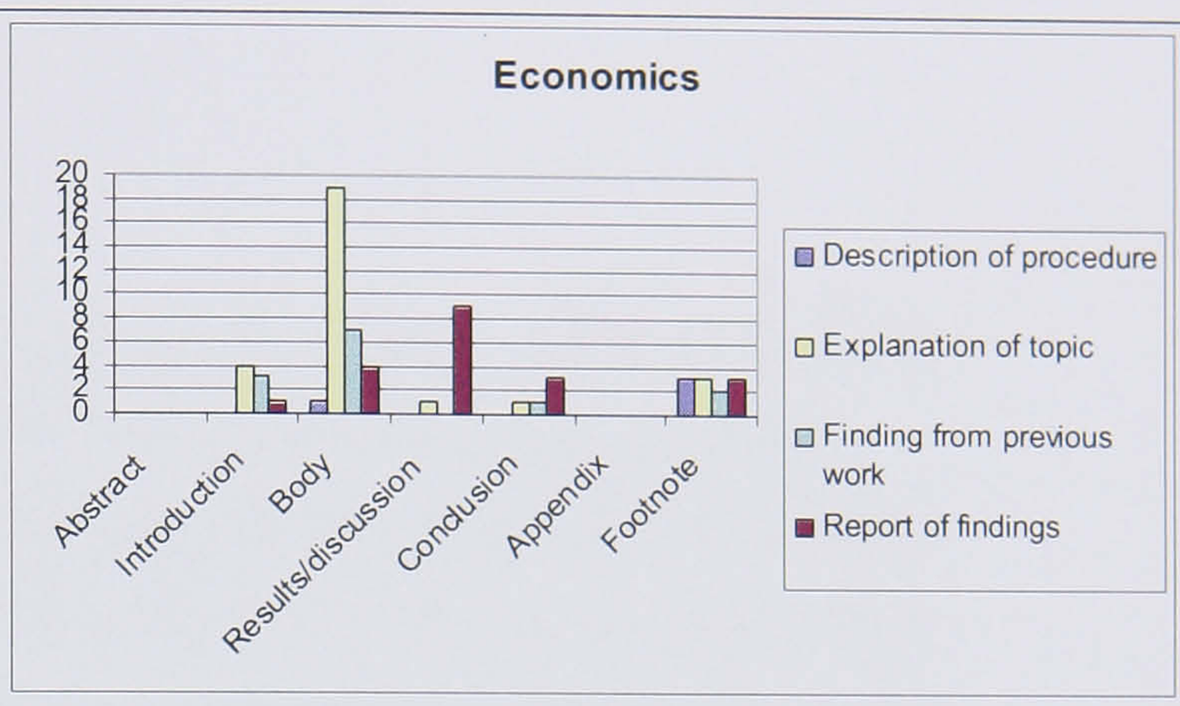
The final relationship to be investigated in this case study is between the discourse function of each isotextual occurrence of the string and its location in the text in which it occurs. The results are shown in Figure 8.8 below, and it can be seen that there is no noticeable distinction between Science and non-Science disciplines with respect to the locations of the sequence with a particular function.

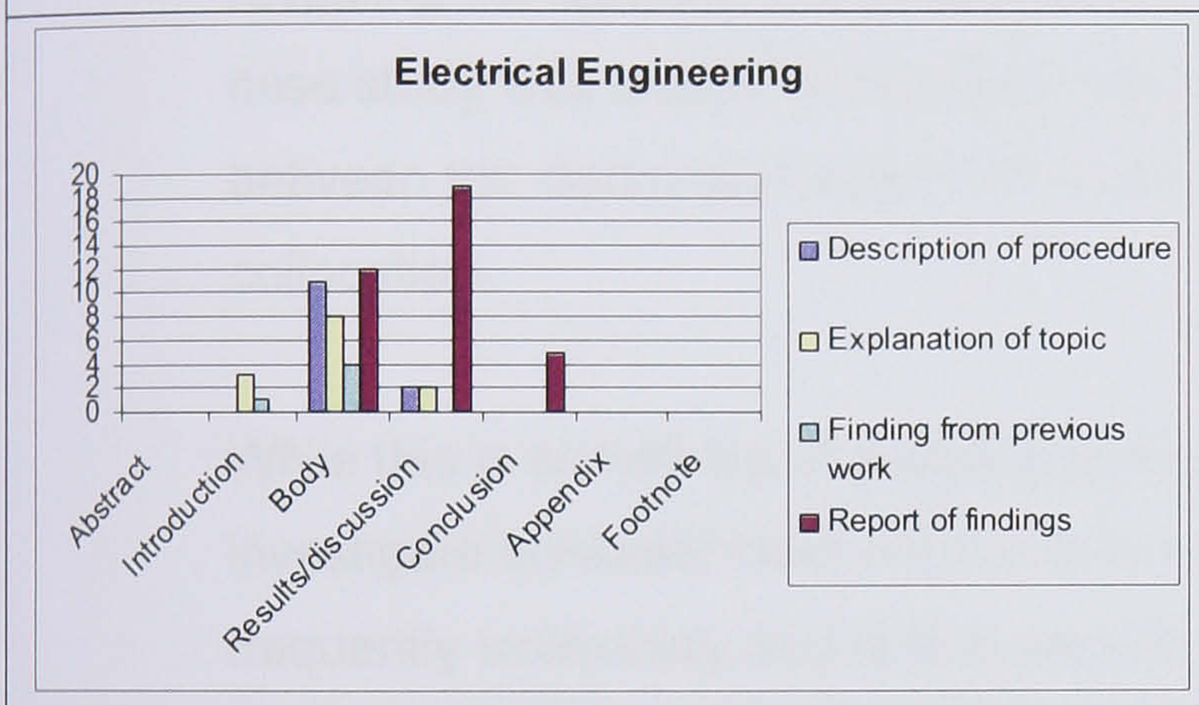
For sequences of *in the case of the* with the function of "finding from previous work", these are found most often in the Body section in Electrical Engineering, Economics, Marketing, and Sociology, in the Results and Discussion section in Cell Biology, in the Introduction section in Mechanical Engineering, and hardly at all in Philosophy. For sequences with the function of "description of procedure", these are found most often in the Body section in Electrical Engineering, Mechanical Engineering, Marketing, and Sociology, again in the Results and Discussion section in Cell Biology, in a footnote in Economics, and again hardly at all in Philosophy.

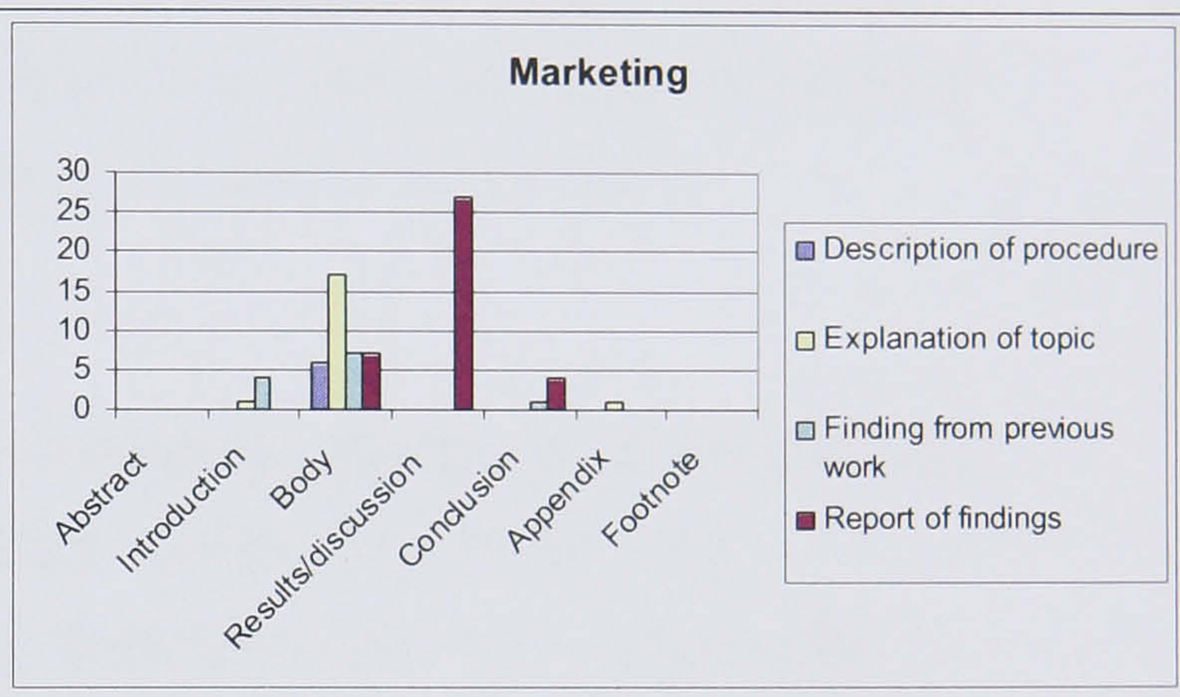
Sequences with the function of "explanation of topic" are found most often in the Body section in all disciplines with the exception of Cell Biology, where they again occur most often in the Results and Discussion section with this function. Finally, when the sequence has the function of "report of findings," it occurs most often in the Results and Discussion section in Cell Biology, Electrical Engineering, Economics, and Marketing, and is equally split between the Body section and the Results and Discussion body in Mechanical Engineering, Nuclear Physics, and Sociology. Once again it does not occur in Philosophy.

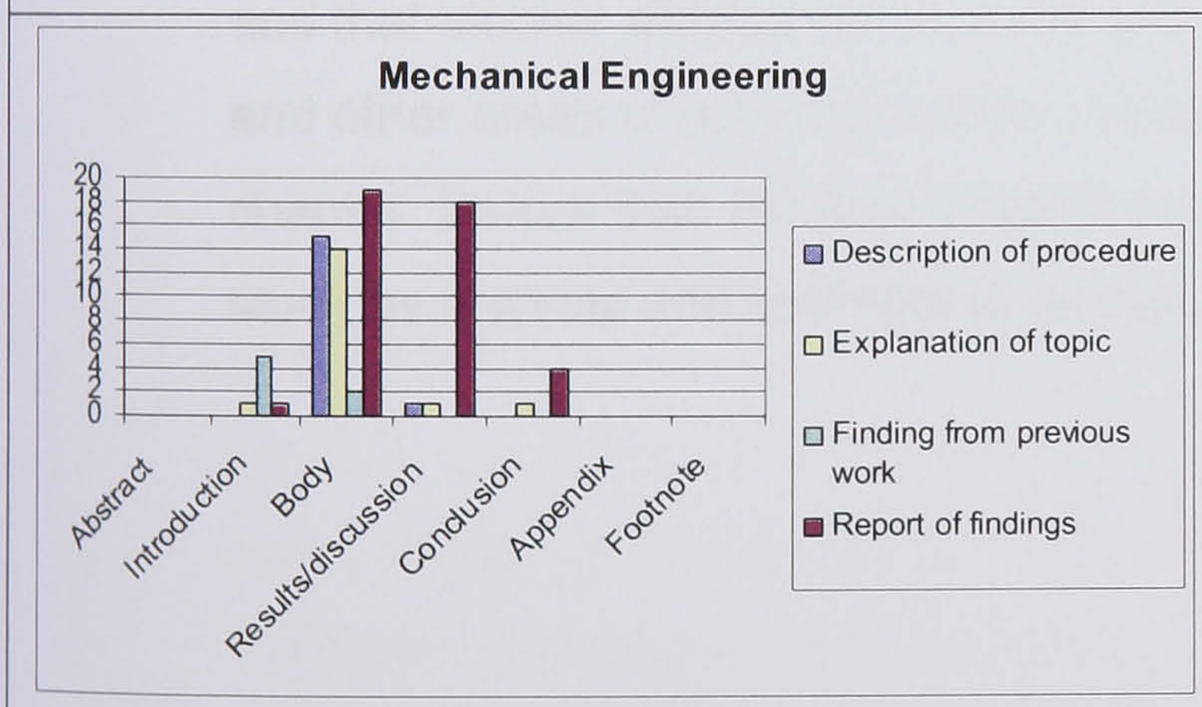
In view of these results, it can be concluded that there is some evidence of a relationship between discourse function and text location for "explanation of topic", and to a lesser extent for "report of findings".

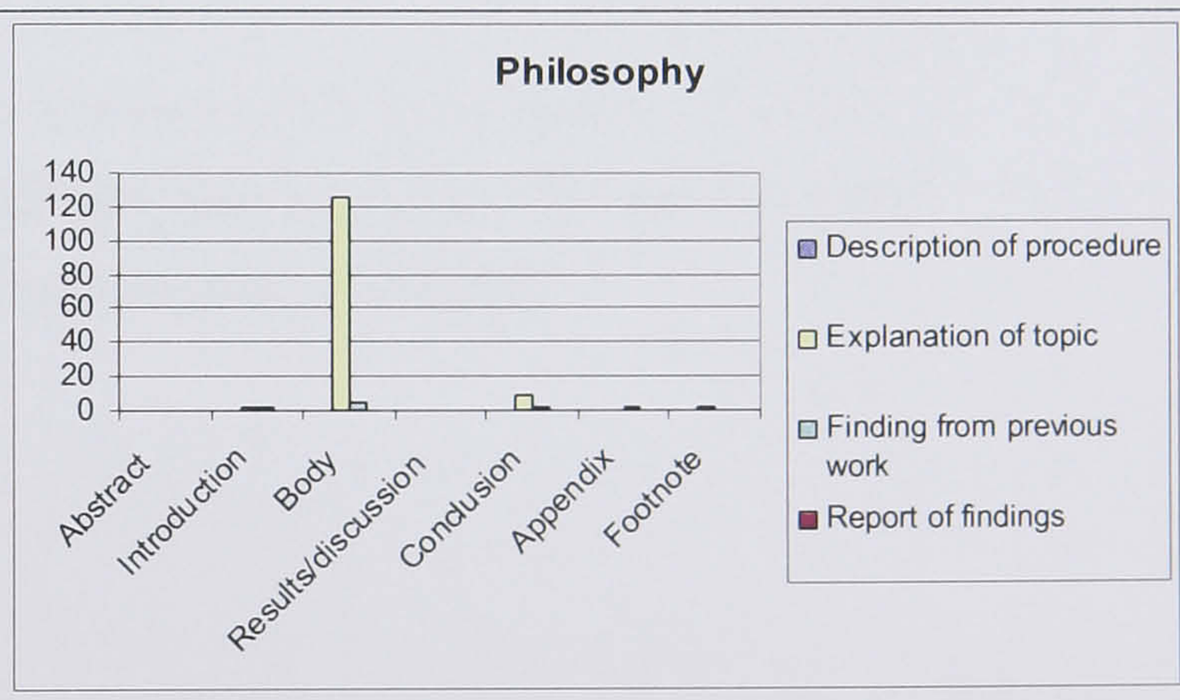


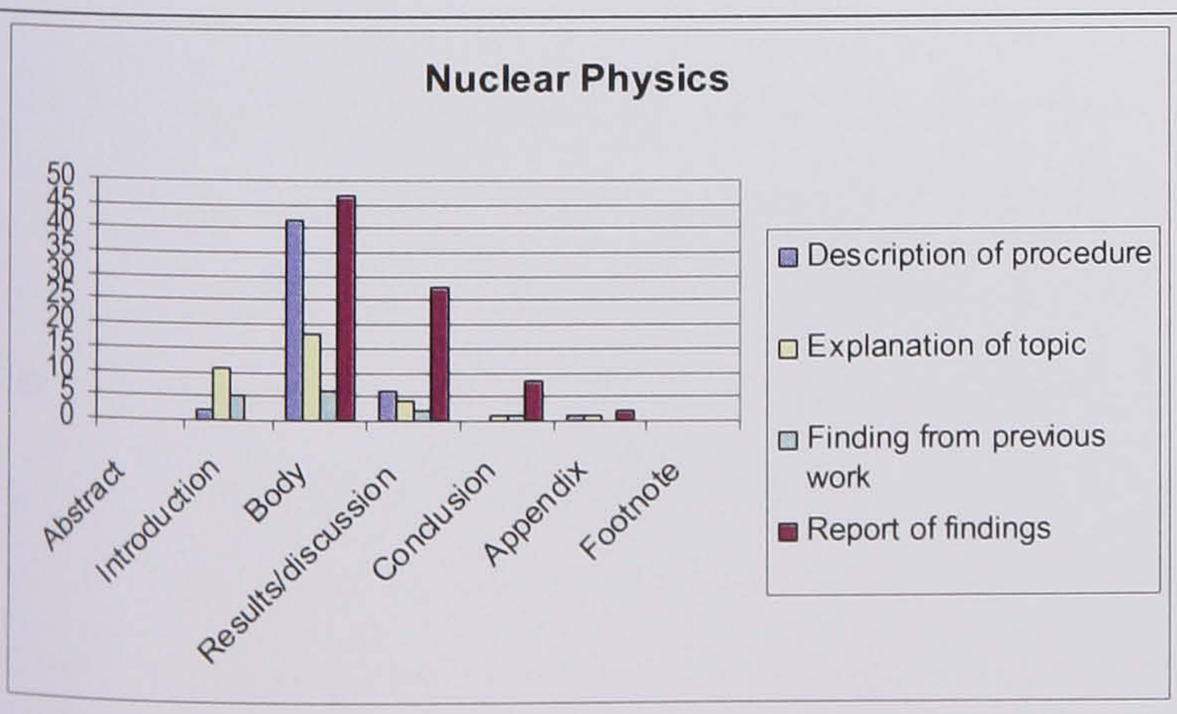












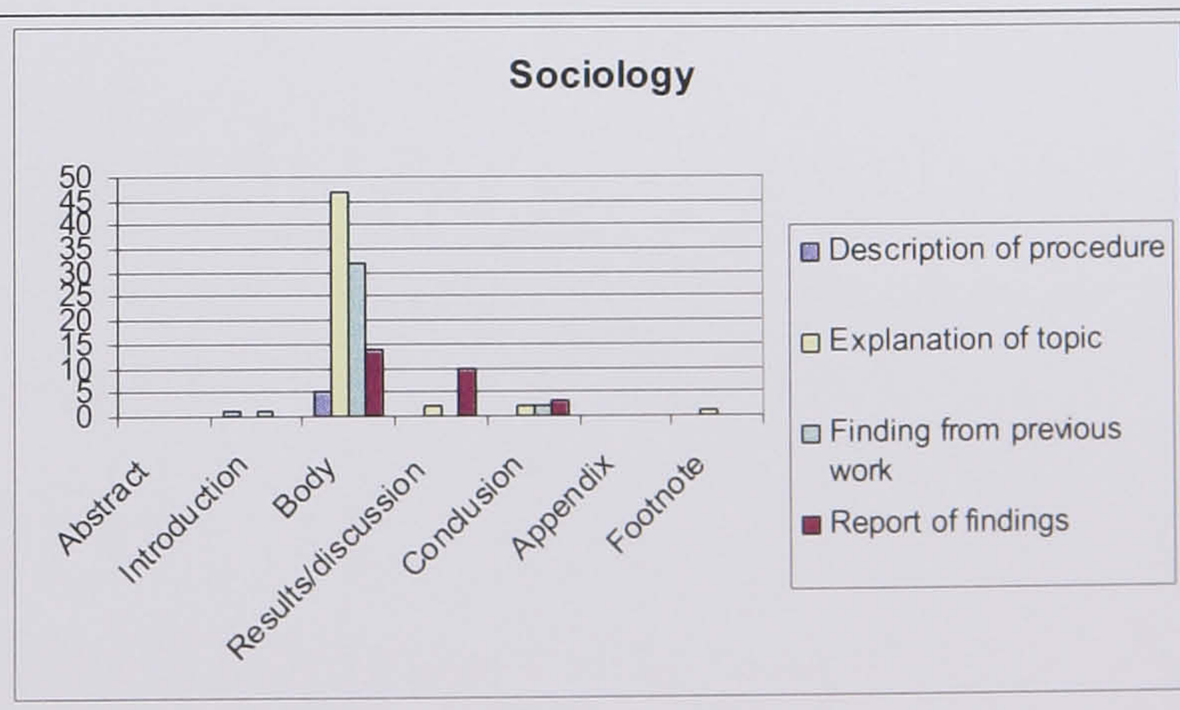


Figure 8.8: The relation between textual location and use of *in the case of the* in each isotextual subcorpus

8.9 Summary

The case study has examined the textual environments and discourse functions of *in the case of* with its most frequent collocate *the*. The lexical, semantic, pragmatic environments of each occurrence were investigated, as well as the location of each occurrence in the text in which it occurred. Although the framework was effective in revealing the different environments of *in the case of the* at the different levels, the case study was unable to discover any particularly noticeable systematic relationship between the discourse function of *in the case of the* and the other levels of collocation.

While this is something of a disappointment, it raises the possibility that attempting to investigate universal fixed collocational patterns, namely those which occur frequently isolexically and isotextually across disciplines, may not be appropriate, and that studies within individual disciplines may be more productive. This possibility, and other areas of potential further research interest, will be discussed in the final chapter. Before that, the next chapter summarises and discusses the results of the study as a whole and attempts to assess its contribution to the field.

Chapter 9 Discussion

9.1 Introduction

This chapter aims to assess the contribution of this thesis to the fields of Phraseology, Academic Discourse, and Corpus Linguistics. It first summarises the key findings of the study in relation to the research questions and discusses the extent to which it was possible to answer them. It then discusses the strengths of this research and its limitations.

9.2 Summary of research findings

9.2.1 Research question 1: isolexical and isotextual fixed collocational patterns

The first research question in this study was concerned with the most frequent fixed collocational patterns in comparable samples of journal articles in different academic disciplines. A contrastive (Sinclair 2005) corpus was constructed in order to answer this question, since it is clear from the overview of existing corpora of academic discourse in Chapter 4 that the component parts of a normative (ibid.) corpus would not have constituted a comparable amount of data. Few available normative corpora which contain journal articles, such as the BNC, have been constructed with comparative studies in mind, and are unsuitable because they include unequal amounts of data in each discipline. In addition, the sampling methodology used in the BNC means that not all texts in each subcorpus are complete, and so the subcorpora do not fully represent all discourse functions. Some stages of the genre of the research article are therefore missing from these subcorpora, and so fixed collocational patterns which may perform discourse functions in these genre stages may also be missing from the results. Other normative corpora such as the BoE contain textbooks rather than journal articles. Each discipline contains an unequal number of texts, and so there is not a like number of genre stages to be compared. Since the number of texts included is rather low, it is also possible for idiosyncratic uses in one or two texts to influence the results.

Answering the first research question, therefore, meant I had to address the issue of what constitutes a comparable amount of data when subcorpora containing complete texts of different lengths are to be compared. Comparing a similar number of tokens in each subcorpus leads to an imbalance in the number of texts in each subcorpus. Comparing a similar number of texts in each subcorpus leads to an imbalance in the number of tokens in each subcorpus. This issue is of particular relevance when comparing fixed collocational patterns with lexical bundles, since the latter have been assigned discourse functions. Comparing subcorpora containing an equal number of tokens but an unequal number of texts risks skewing the results in favour of the subcorpus containing the most texts. If a pattern is found to occur more in a subcorpus containing many short texts than in a subcorpus containing a few long texts, then it can be difficult to determine whether there is a tendency for the pattern to be used more often in that discipline than others, or whether the pattern merely appears frequently because of the imbalance in the number of texts in that discipline. The relationship between a pattern's discourse function in texts and its frequency of occurrence in the subcorpus for that discipline is therefore difficult to discern using an isolexical comparison.

The issue of comparability was addressed by constructing two comparable versions of the corpus - one comparable in terms of texts, the other comparable in terms of tokens - and so a distinction between isolexical and isotextual corpora was introduced. The first question aimed to identify the most frequent fixed collocational patterns in both an isolexical sample containing an equal number of tokens in each discipline, and an isotextual sample containing an equal number of texts from each discipline. The isolexical measure chosen was one million tokens per journal, and the isotextual measure chosen was 100 texts per journal, the number required to collect one million tokens in Marketing. If any pattern occurred more frequently in a discipline than other disciplines, the imbalance from text numbers on the results could be discounted.

Several differences between the isolexical and isotextual versions of the corpus were noted. The most obvious was the difference in size (measured in tokens) caused by the difference in text lengths between disciplines. In non-Science disciplines, the

isotextual subcorpora tended to be as large as or larger than the isolexical subcorpora, whereas in Science disciplines the isotextual versions were between 80% and 33% of the size of the isolexical versions. A version size differential (VSD) measure was developed to indicate the difference in size. The study focused on 3- and 4-word patterns since there were far fewer patterns greater in length which were found to occur with the minimum frequency. In overall terms the difference in the overall numbers of pattern types and tokens between isolexical and isotextual versions varies between Science and non-Science disciplines. In Science disciplines the difference between the number of isolexical and isotextual occurrences is larger than the VSD for those disciplines, while in non-Science disciplines the difference between isolexical and isotextual type and token frequencies is largely the same as the VSD.

In order to answer the first research question, two sets of results, one isolexical and one isotextual, were obtained for 3- and 4-word patterns in each discipline. The results were ranked in order of frequency and then the two frequency distributions were compared and differences in rank positions calculated. This comparison of rankings produced the most surprising finding of this study as a whole. The answers for question 1a) and 1b) are to all intents and purposes the same, as the rank differences for the top 50 most frequent patterns are very low. This means that the most frequent fixed collocational patterns in an isolexical sample of journal articles in different academic disciplines are also the most frequent fixed collocational patterns in an isotextual sample of journal articles in different academic disciplines. This finding is surprising because it holds true in every discipline, regardless of Marketing's VSD of 1.00 or Electrical Engineering's 0.34. This study has shown, therefore, that in individual disciplines frequent fixed collocational patterns are frequent irrespective of the length of research articles in that discipline.

9.2.2 Research question 2: isolexical and isotextual fixed collocational patterns across disciplines

The second research question concerned whether fixed collocational patterns occurred across all eight disciplines. A relatively small percentage of isolexically frequent 3-word patterns were found to occur in each discipline (9% or less), and a slightly higher percentage of 3-word patterns were found to occur isotextually (15% or less). Even smaller proportions of 4-word fixed collocational patterns were found to occur in all eight disciplines. The results for 4-word patterns were even lower: 4% of isolexically frequent 4-word patterns or less were found to occur in each discipline, and 10% of 4-word patterns or less were found to occur isotextually. This is similar to Hyland's finding that for the top 50 4-word bundles in each discipline only five occur in all four disciplines (Hyland 2008: 13).

The findings in relation to the second research question are therefore fairly clear: the different disciplines have few 3-word fixed collocational patterns in common, and even fewer 4-word patterns in common. Both isolexical and isotextual subcorpora in non-Science disciplines share more 3- and 4-word patterns than Science disciplines, a finding that would not have been confirmed by a purely isolexical comparison. The rank differences between isolexical and isotextual subcorpora for those fixed collocational patterns which occur in all eight disciplines are wider than the rank differences for fixed collocational patterns in single disciplines, although not by much: around 75% of the patterns across disciplines still have a rank difference of 10 or below.

9.2.3 Research question 3: comparability

The third research question concerned the comparability of the isotextual fixed collocational patterns from the present study with results from other studies of isolexical fixed collocational patterns, namely the lexical bundles found by Biber (2006) and Hyland (2008). The fact that all three studies are of different levels of discourse: the textbook genre (Biber 2006), the register of academic prose consisting of research articles, doctorates, and MA dissertations (Hyland 2008), and the research article genre (the present study), might suggest that it is not useful to

compare occurrences of isolexical fixed collocational patterns (i.e. lexical bundles). However, the finding from the first research question in this study, namely that frequent isolexical fixed collocational patterns are also frequent isotextually, needed to be tested against isolexical studies of genre and register corpora. The isotextual fixed collocational patterns from the present study were therefore compared with the lexical bundles in Biber and Hyland.

The frequent lexical bundles in both studies, and the discourse functions ascribed to them, were tabulated and compared with the isotextual fixed collocational patterns in this study. There was little overlap between the results from Biber's study and the results from Hyland's study. This lack of overlap reveals that there is a limited number of 4-word lexical bundles which are common across academic textbooks and the register of academic prose. It is important to determine, therefore, whether those lexical bundles which *are* common also occur as isotextually frequent fixed collocational patterns in the present study. It was found that this was indeed the case: *on the other hand*, *as well as the*, and *in the case of*, which are frequent in all disciplines in both Hyland and Biber are also frequent in every discipline in the present study. None of these patterns are assigned a stance or interpersonal discourse function by Biber or Hyland's framework: they are primarily text organisers, although Biber classes *in the case of* as a referential bundle. Several other "missing" fixed collocational patterns were identified, namely those occurring isotextually in more than one discipline in the present study but which are not listed among Biber's or Hyland's lexical bundles.

The results of the comparison between the isotextual fixed collocational patterns in the research article genre in the present study and the lexical bundles in the textbook genre in Biber (2006) and the lexical bundles in the academic register of Hyland (2008) are therefore inconclusive. Only three frequent isolexical fixed collocational patterns (i.e. lexical bundles) also occur as frequent isotextual fixed collocational patterns in the present study. This may be because the lexical bundles in both Biber's and Hyland's studies were normed upwards to obtain an isolexical "per million words" measure of frequency, and so the subcorpora in Biber's study and Hyland's study, which fewer contain than a million words, are too small, and are not

therefore comparable with the data in the present study. The threshold for similarities between lexical bundles in academic genres and registers may lie above this level.

The forms of isotextual 4-word fixed collocational patterns were also compared with the forms of the 4-word lexical bundles in Hyland's study. Broad agreement was found in four structural categories between the percentages of fixed collocational patterns in the present study and the percentages of the lexical bundles in Hyland's study (2008). Only in patterns of the formal categories of *noun phrase + of-fragment* and *other* was there no noticeable similarity between the two studies.

Finally a case study was made of the only frequent isotextual fixed collocational pattern to occur in the present study, in the two lexical bundle studies of Biber and Hyland, and in the study of academic formulas by Ellis and Simpson. This aimed to investigate the collocational environment of this pattern in relation to the four levels of collocation used in "Eastern European" Phraseology and in the "Empirical Firthian" tradition. The heuristic framework incorporating these levels seemed to provide a useful tool for investigating the environments of this pattern, but yielded few interesting results for *in the case of the*. It was not possible to find any common semantic environment by examining the frequent head nouns framed by this sequence. Four discourse functions were identified for the sequence, two of which, "explanation of topic" and "report of findings" had an observable relationship with its location in text. It may be that an investigation of one pattern across eight disciplines was too ambitious, and that more useful results might be obtained from studies within a single discipline.

Having summarised the answers to the research questions, I now move on to evaluating the strengths and weaknesses of how the study tackled them in this thesis.

9.3 Strengths of the research

9.3.1 Corpus construction and methodology

The corpus constructed for this study is large in relation to corpora collected for previous studies of phraseological features in academic discourse. A considerable amount of care and attention has been paid to its construction in order to ensure that it contains comparable subcorpora of texts from prestigious journals in each discipline. Both versions of these subcorpora are also large in relation to previous studies, meaning that many occurrences of both isolexical and isotextual patterns could be found. The numbers of occurrences are consequently larger than in previous studies, meaning that the frequencies of fixed collocational patterns stand as they are and do not need to be normed or otherwise adjusted. This lends the results of this study a robustness not found in other studies with normed frequencies.

The size of the corpus also adds to the strengths of the claims that can be made for these findings in relation to the genre of the research article. Previous influential studies, such as those on research article introductions, have made strong claims based on detailed analysis of a limited number of texts. The case study in the previous chapter, on the other hand, looked at one pattern in 493 texts. *In the case of* was not simply chosen, moreover, as an item of interest for further study: it chose itself as the best possible candidate by emerging as the most frequent pattern in the data in this study and three others.

9.3.2 Breadth of scope

This study is broad in scope, and offers insights into the Phraseology of language, into Genre and Register, and into Corpus Linguistics. It has required a detailed understanding of the numerical, technical, and statistical issues in corpus compilation. At the same time, it has also been important to keep the focus of the investigation on the fact that the data in a corpus is the instantiation of communicative acts in particular discourse communities, rather than merely a large amount of language which can be abstracted into numbers and treated in the same way as other numerical data. This has required as much understanding of

collocational theory and the nature of language, and the applications of skills in the analysis of staged, situated discourse, as a non-corpus-informed study would have required.

The study adds to knowledge of Phraseology by revealing more about the forms, functions, and textual environments of the fixed collocational pattern as a phraseological unit, and by proposing a framework for which these collocational environments can be further investigated; it adds to knowledge of the genre of the research article by revealing more about the relationship between the location and discourse function of fixed collocational patterns; it adds to knowledge of Corpus Linguistics by proposing the distinction between isolexical and isotextual comparative corpora, and by suggesting the possibility of a threshold above which the rankings of fixed collocational patterns may be stable.

9.4 Limitations of the research

9.4.1 Data and software

One possible area which might limit this study is the amount of noise in the data. For the reasons described in section 5.3, extra-textual noise was included in the data to allow time for more texts to be collected and for the corpus to be consequently larger. Estimates were made of the amount of noise likely to be in the texts in each subcorpus, and allowance was made for this by collecting more than a million tokens per journal in the isolexical version, so that at least a million running words were included. The problem that might remain is that some of the estimates for noise were not accurate, resulting in over- or underestimation of the actual number of running words in each subcorpus. This would be unlikely to affect the frequency counts for fixed collocational patterns, since noise is extracted before the pattern frequency lists are compiled, as explained in section 5.5.1, but it could have affected the calculation of the version size differential (VSD).

The reason for this is that if texts in a discipline contain more noise than allowed for, then both its isolexical subcorpus and its isotextual subcorpus will contain fewer running words than realised. For a discipline like Marketing, where the isolexical and

isotextual versions both contain similar numbers of texts, the version size differential will still give a fair reflection of the difference in size of the two versions because the number of tokens in each version will be similar: the VSD will remain near 1.0 despite the underestimation of noise. In disciplines such as Electrical Engineering, however, where the isolexical subcorpus contains many more texts than the 500 in the isotextual subcorpus, there is disproportionately more noise in the isolexical subcorpus than in the isotextual subcorpus. This extra noise could artificially inflate the size of the isolexical subcorpus and make the VSD lower than if the noise estimates were accurate.

One other type of noise, termed in this study as "fluff", also has the potential to affect some of the rankings of any fixed collocational patterns containing ligatures such as *fl* and *ff*. If the corpus had been collected slightly later, text could have been stored in Unicode rather than ANSI format. The Unicode system contains characters representing these ligatures, and so less fluff would have been included had Unicode been available at the time the corpus was collected. This limitation is an almost inevitable consequence of the rapid technological developments in any area of computing. In the case of software, the fully functional shelf-life of a product such as *WordSmith Tools* is under three years: *WordSmith Tools* version 4 used in this study was released in 2005 and continually updated with bug fixes throughout the two years during which I was measuring the size of the corpus, identifying patterns, and analysing their discourse functions. *WordSmith Tools 5* was released in Spring 2008, by which time this thesis was nearly completed; it contains additional clustering functions which may reveal other aspects of fixed collocational patterns but which cannot now be considered in this study.

An additional limitation is that other lexical processing software, such as Wmatrix (Rayson 2008), SketchEngine (Kilgarriff et al. 2004), and ConcGram (Greaves 2005), have been released since I began analysing the corpus, and this study might have found other insights had I been able to use them. Also, other large-scale corpora of academic research articles in different disciplines, particularly the Brigham Young University Corpus of American English (Davies 2008), have recently been released and may have been comparable for this study.

9.4.2 Statistical significance

Another possible limitation is the lack of any statistical measure of the significance of the frequencies of fixed collocational patterns in each version of the corpus. The reasons why this study relied on raw frequency counts instead of statistical measure were discussed in section 5.5.3 above, but it remains that some corpus linguists and researchers in Phraseology may view this reliance on raw frequency as a limitation of the study.

9.4.3 Intuitive functional categories

A further limitation of the analysis in the case study is my reliance on intuition to group the occurrences of *in the case of the* into functional categories. Such a subjective methodological approach would seem at odds with the rigorous empirical corpus-driven approach originally taken in identifying this pattern as one which merited further study. And yet, as was argued in 9.3.2, one of the strengths of this study is that it employs both a corpus-driven approach and a traditional discourse analytical approach. Subjective meaning in language cannot be automatically analysed by a computer without prior annotation of the data by a corpus, or "training" of the analysis software on other annotated data. The computer is merely a tool for arriving at the linguistic feature to be studied, which must then be analysed by a human linguist.

9.5 Summary

This chapter has summarised how this study has answered the research questions, and evaluated the strengths and weaknesses of the way in which they were answered. We now move on to the final chapter, which discusses some of the implications of the study and the opportunities it creates for further research.

Chapter 10 Implications and further research

10.1 Introduction

This final chapter discusses the implications of this study for the field of comparative corpus studies of Phraseology in Academic Discourse. It suggests areas where further research might clarify some of the remaining issues in Corpus Linguistics, Academic Discourse, Phraseology, and finally Psycholinguistics and Pedagogy.

10.2 Corpus Linguistics

The results of the present study are likely to have considerable implications for researchers working on studies of comparative corpora. Other researchers, however, may feel that the results do not require any reconsideration of what constitutes a comparable amount of data. It might be thought that the similarity identified in this study between the isolexical and isotextual rankings of frequent fixed collocational patterns means that there is actually no need for a distinction to be made between isolexical and isotextual data sets in comparative studies of fixed collocational patterns. If there is little difference between the frequency rankings for frequent patterns in the two versions, then quantitative findings of comparisons between isolexical corpora are perfectly valid, and the imbalance of texts between subcorpora will not bias the results.

I would instead argue that this study has only established that there is a threshold above which this similarity exists, and it is clear that neither existing comparative corpora, nor previous studies of fixed collocational patterns, contain enough texts or tokens to exceed this threshold. Other studies will therefore need to take account of the findings of this study unless each of their comparative corpora consist of at least 5 million tokens or 500 texts, or the results of anticipated further research (see below) establish a lower threshold. The important isolexical/isotextual distinction would not have been drawn if the present study had not been undertaken, and it should encourage more corpus linguists to think of their data in terms of the amount

of communication it contains rather simply in terms of the quantity of language it contains.

Other comparative studies of the discourse functions of fixed collocational patterns in a genre, whether isolexical or isotextual, should also consider the implications of this distinction. When the discourse functions of a fixed collocational pattern are being investigated I would argue that an isotextual comparison is more useful than an isolexical comparison for comparative studies of the same genre. As I have argued, discourse functions are related to stages in a genre, and so an isotextual measure reveals more about the relationship between form and function.

There is one other issue raised by the introduction of the construct of the isotextual fixed collocational pattern, namely the ontological status of the lexical bundle as an isolexical form-function construct. Lexical bundles are defined isolexically in terms of minimum occurrences per million words, but also are assigned discourse functions, which are features of genre stages in text. The present study, however, has suggested that isotextual studies of fixed collocational patterns give a clearer indication of the discourse functions in a genre. This finding in turn suggests that as isolexical constructs, lexical bundles cannot reliably be assigned discourse functions, and are therefore not as useful as first thought.

While the similarity between isolexical and isotextual rankings of frequent fixed collocational patterns has been highlighted by this study, it remains to be satisfactorily explained. Further work would need to compare different isolexical and isotextual versions of the same corpus. The different versions of the isolexical corpus could be constructed, perhaps in progressively smaller sizes from 800,000 tokens down to 200,000 tokens per journal, so that the relationship could be tested in subcorpora containing from four million to one million tokens. Comparisons could also be made between isotextual corpora containing the oldest 500 texts rather than the 500 most recent texts, and the number of texts could also be reduced progressively from 80 to 20 to see how far the isolexical/isotextual relationship holds. These further studies would help establish whether a threshold exists below which the isolexical/isotextual similarity is no longer discernable.

10.3 Academic Discourse

The field of Academic Discourse would also benefit from further study of the isolexical/isotextual distinction for a number of reasons. First, this relationship has only been established in research articles in eight disciplines. It would be important to see whether it still pertains in research articles in other disciplines, such as Law or Medicine. Second, it would also be important to learn whether this relationship is a feature of other written academic genres, such as student dissertations or textbooks. While this study has shown that a limited number of the isotextual fixed collocational patterns in this study were also frequent in the textbook studies by Biber (2006) and the study of dissertations by Hyland (2008), their corpora were not designed for the isolexical/isotextual distinction to be tested. The obvious difficulty here would be to collect progressively larger numbers of textbooks in order to establish a large enough isotextual corpus.

In view of the weak relationship between the discourse function of *in the case of the* and its syntactic, semantic, and textual environments across disciplines in the case study, it would seem that a more fruitful area of investigation of functions of isotextual fixed collocational patterns would be within single disciplines. Using the framework employed in the case study, isotextually frequent fixed collocational patterns which are frequent in one discipline could be examined. This would reveal more about the context of situation of each discipline through its lexical environment, and also show whether the relationship between the pattern's discourse function and its textual location is any stronger.

The isolexical/isotextual distinction could be further investigated beyond academic discourse by constructing isolexical and isotextual corpora of texts in other genres. Recent work by O'Donnell (2007) has clarified the relationship between the lexical priming of a word and its textual location using the *Guardian News Archive,* a large corpus of 60,000 newspaper articles. It would be a simple matter to collect different 500-text and one-million-word samples from this corpus and compare the resultant isolexical and isotextual fixed collocational patterns. The final area where the results of this study could be investigated further is the opinion of the members of the discourse community. It would be useful to discuss the frequent fixed collocational

patterns with academics from the relevant disciplines and gain a clearer idea of the functions they assign to them.

10.4 Phraseology

The results from the present study also have a number of implications for work on Phraseology. It would be interesting to test the isolexical/isotextual distinction on the phenomenon of collocation in general. As noted in section 2.3, the Empirical Firthian tradition identified collocational relations between lexical items based on corpus evidence, but the corpora used were normative, and the context of situation was taken as its co-text in a span of 5 words to the left or right. It would be interesting to see whether collocation has isolexical and isotextual dimensions: are the frequent collocations of a word the same isolexically as they are isotextually? Such studies would reveal more about the semantic behaviour of individual words, not merely how they are used, but how they are used in a particular number of communicative situations.

Another issue for further investigation is the possibility of a collocational "span" around a fixed collocational pattern. The results of the case study suggest that this would probably be best attempted on different patterns in the same discipline rather than for one pattern across disciplines. One pattern may have one set of collocates isolexically and another set isotextually, and then be found to have completely different sets isolexically and isotextually in another discipline.

Isotextual fixed collocational patterns would also be worthy of further investigation in the wider area of lexico-grammar. Previous work on pattern grammar by Hunston and Francis sees lexical phrases and other phraseological form-function constructs as rather peripheral items:

> collections of lexical phrases are, ultimately, fairly random lists of phrases, organised either according to their relative fixedness, or to their function (discourse organising, opinion giving and so on) or to one of their core words... They are an attempt to account for a portion only of the lexicon. Grammar patterns, on the other hand, constitute an attempt to describe the whole of the language (or rather, all the frequently-occurring items in the language) in a principled way, and the lists of words collected in a given pattern are not random.
>
> (Hunston & Francis 2000: 14)

The results found in section 7.2, however, were that the forms of lexical bundles identified in previous studies are largely similar to the forms of the fixed collocational patterns in the present study. It may be that the forms abstracted from the patterns found in the discourse, such as NP+OF, and IT+AP/VP, could be seen as a link between the grammar patterns of Hunston and Francis and the lexical phrases they criticise. The isotextual nature of these fixed collocational patterns connects their form more securely to their function, and the stability of their isolexical/isotextual frequency rankings observed in all eight disciplines means that they are by no means random.

10.5 Psycholinguistics and Pedagogy

In addition to the study of language on paper or in the air (Sinclair 2001: ix), the final area for which this study has implications is in the study of language in the mind, namely in investigations of the role of corpus-derived forms in cognition and language learning. Early work on attempting to test whether frequent fixed collocational patterns identified from corpora are also stored holistically and used in language production was attempted by Schmitt et al. (2004: 128), who tested the hypothesis that "the language in corpora has been produced by people using language and so must reflect language competence to some extent" (2004: 128). Their findings suggest that "corpus data on its own is a poor indicator of whether these clusters are actually stored in the mind as wholes" (ibid.). The study by Schmitt et al. has a number of limitations which suggest that its results are not conclusive, and that a closer attention to the context of situation of the target patterns used in such experiments may produce clearer results. These limitations will be briefly summarised here.

The first issue is with the target corpus-derived fixed collocational patterns which they intended to use as an indicator that patterns were stored holistically in the minds of the subjects. These are shown in Table 10.1 below. These patterns cannot be considered to be completely corpus-derived, however. Schmitt et al. drew up the above list from existing work on Phraseology, much of which has already been discussed in Chapter 2.

aim of this study	*in the same way as*
as a consequence of	*is one of the most*
as a matter of fact	*it was going to*
as shown in figure	*it's not too bad*
for example	*night and day*
from the point of view	*on and off*
go away	*something like that*
I don't know what to do	*to give you an example*
I see what you	*to make a long story short*
in a variety of	*what I want to*
in addition to the	*you know*
in the middle of the	*you've got to have*
in the number of	

Table 10.1: Fixed collocational patterns in Schmitt et al. (2004: 130)

These patterns were taken from the lexical bundles in Biber et al. (1999), the lexical phrases of Nattinger and DeCarrico (1992), and others were also found from corpus searches for some of Hyland's words expressing doubt and uncertainty (Hyland 2000: 188). Biber et al's lexical bundles, as we have seen, were identified in register corpora and so are corpus-derived, although they are found in two very different registers. Nattinger and DeCarrico's lexical phrases, as we have also seen in section 4.4.1, were taken from a vaguely defined data set which cannot, in view of the lack of information regarding its construction and content, be regarded as a corpus in the usual sense of the term. It is also less certain that the patterns derived from Hyland's work are also corpus-derived. Hyland's original list of words was not identified from corpus data but was instead taken from the existing non-corpus-based literature on evaluation (Hyland 2000: 90). It is difficult to regard patterns based on corpus searches for these words in the same way as empirically-identified fixed collocational patterns.

The second issue is with the methodology used by Schmitt et al. A story was written with these patterns "embedded" (2004: 131) in it and which was dictated to subjects in bursts. The subjects then were recorded retelling the story and the number of patterns in their recordings counted. From Table 10.1 it would seem that these patterns come from many different registers, and possibly from more than one genre within each register. I would argue that the existence of patterns from such disparate contexts of situation, e.g. *as shown in figure* and *it's not too bad,* embedded in one genre, the story, creates a highly unusual text (ibid.: 149-151). Schmitt et al's finding

that few subjects used these patterns when retelling the story could be the result of this dislocation between the pattern role in the story (i.e. its artificial context of situation) and the actual genre from it was taken, rather than the result of any factors relating to cognition.

I would suggest that any further work arising from the present study which goes on to investigate the role of fixed collocational patterns in language cognition and production should not treat patterns as being independent of their context of situation. An attempt could be made to try and replicate Schmitt et al's study while using more rigorously corpus-derived fixed collocational patterns. The criterion of genre compatibility should also be applied so that patterns identified as target patterns are presented in a sample text from the same genre, and possibly the same contexts of situation as that in which the target patterns were found.

In addition to further studies on the role of fixed collocational patterns in language cognition and production, there could also be more investigation of whether the use of fixed collocational patterns might be pedagogically suitable. Although several attempts have been made to base EAP teaching materials on findings from corpus data, e.g. (Oakey 2002b; Harwood & Hadley 2004; Harwood 2005), a pessimistic view is that of Swales (2002: 152), who asserts that much of the "enthusiastic yet fragmented efforts" to identify pedagogically useful features from corpus studies can be regarded "as banked intellectual resources whose pedagogical time has yet to come" (ibid.: 159). However, if fixed collocational patterns are found to have a role in language cognition, they might be more pedagogically relevant than other features identified from corpus data.

The most appropriate pedagogical area to investigate would be determining the role of fixed collocational patterns in improving the accuracy of students' lexico-grammar. In my experience, students' inaccuracies in production seem to stem less from difficulties with unfamiliar words and more from problems with unfamiliar combinations of familiar words (Oakey 2005). Fixed collocational patterns can be presented to learners as lexical words frozen in the act of combining with their accompanying grammatical words, such as *in the regulation of*, *have been shown to*, and *that it is a*. They present an alternative pedagogical unit to the word or clause or

sentence, and could be used, in a complementary way to the *COBUILD Grammar Patterns*, for raising students' awareness of the lexico-grammatical environments of the vocabulary they are likely to need to use.

10.6 Conclusion

Having discussed the many possible areas for future research arising from the findings of the present study, it is now time for me to draw this thesis to a close. The original aim of the study was to make a comparative study of the forms and functions of fixed collocational patterns in a subset of written academic English, and in attempting to achieve this goal the isolexical/isotextual similarity was discovered. The study has made a number of other findings which clarify our understanding of fixed collocational patterns in the genre of the research article, and has suggested a method of finding out more about the discourse functions of fixed collocational patterns in texts. It has also raised a number of questions in relation to Phraseology, Academic Discourse, and Corpus Linguistics. This final chapter has outlined the agenda for future research, an agenda which I very much look forward to pursuing.

References

Aisenstadt, E. (1981). Restricted Collocations in English Lexicology and Lexicography. *ITL (Instituut voor Toegepaste Linguistiek) Review of Applied Linguistics, 53*, 53-61.

Aston, G. (2001). Text Categories and Corpus Users: A Response to David Lee. *Language Learning and Technology, 5*(3), 73-76.

Barber, C. L. (1962/1985). Some Measurable Characteristics of Modern Scientific Prose. In J. M. Swales (Ed.), *Episodes in ESP*. Oxford: Pergamon.

Barlow, M. (2004a). *Collocate*. Houston: Athelstan.

Barlow, M. (2004b). Software for Corpus Access and Analysis. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 205-221). Amsterdam: John Benjamins.

Becker, J. D. (1975). The Phrasal Lexicon. In B. Nash-Webber & R. Schank (Eds.), *Theoretical Issues in Natural Language Processing 1* (pp. 70-77). Cambridge, Mass.: Bolt, Beranek, and Newman.

Benson, M., Benson, E., & Ilson, R. (Eds.). (1997). *The BBI Dictionary of English Word Combinations* Amsterdam: John Benjamins.

Bhatia, V. K. (1993). *Analysing Genre: Language in Use in Professional Settings*. London: Longman.

Biber, D. (1988). *Variation in Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University English: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes, 26*, 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If You Look At ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing Language in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton: Educational Testing Service.

Biber, D., Johansson, S., Leech, G. N., Conrad, S. M., & Finegan, E. (Eds.). (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bloor, M., & Bloor, T. (1989). Cultural Expectations and Socio-Pragmatic Failure in Academic Writing. In P. Adams, B. Heaton & P. A. Howarth (Eds.), *Socio-Cultural Issues in English for Academic Purposes* (pp. 1-12). London: MacMillan.

Bolinger, D. (1976). Meaning and Memory. *Forum Linguisticum, 1*(1), 1-14.

Brier, N. (2005). Untitled Discussion List Posting. from http://longtail.typepad.com/the_long_tail/2005/01/definitions_fin.html

Butterfield, J. (Ed.). (2003). *Collins English Dictionary 6th Edition*. Glasgow: HarperCollins.

Canagarajah, S. (1999). *Resisting Linguistic Imperialism in English Teaching*. Oxford: Oxford University Press.

Carter, R. (2004). *Language and Creativity: The Art of Common Talk*. London: Routledge.

Carter, R., & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Charles, M. (2004). *The Construction of Stance: A Corpus-Based Investigation of Two Contrasting Disciplines.* Unpublished PhD, University of Birmingham, Birmingham.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics, 16*(1), 22-29.

COBUILD. (1995). *COBUILD English Collocations on CD ROM*. Glasgow: HarperCollins.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., & Ferrera, J. (1988). Reading English for Specialized Purposes: Discourse Analysis and the Use of Student Informants. In P. L. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive Approaches to Second Language Reading*. Cambridge: Cambridge University Press.

Cortes, V. (2002). Lexical Bundles in Freshman Composition. In R. Reppen, Fitzmaurice, Susan, Biber, Douglas (Ed.), *Using Corpora to Explore Linguistic Variation* (pp. 131-145). Amsterdam: John Benjamins.

Cortes, V. (2004). Lexical Bundles in Published and Student Writing in History and Biology. *English for Specific Purposes, 23*(4), 397-423.

Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Application.* (pp. 1-20). Oxford: Clarendon Press.

Cowie, A. P. (1998b). *Phraseology: Theory, Analysis, and Application.* Oxford: Clarendon Press.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Crystal, D. (1996). *Rediscover Grammar*. London: Longman.

Crystal, D. (2003). *English as a Global Language* (2 ed.). Cambridge: Cambridge University Press.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-Aided Error Analysis. *System, 26*, 163-174.

Danielsson, P. (2007). What Constitutes a Unit of Analysis in Language? *Linguistik Online, 31*, 17-24.

Dillon, G. L. (2006). Corpus, Creativity, Cliché: Where Statistics Meet Aesthetics. *Journal of Literary Semantics, 35*(2), 97-103.

Dudley-Evans, A. (1986). Genre Analysis: An Investigation of the Introduction and Discussion Sections of M.Sc. Dissertations. In M. Coulthard (Ed.), *Talking About Text*. Birmingham: English language Research, University of Birmingham Press.

Dudley-Evans, A., & St John, M.-J. (1998). *Developments in ESP*. Cambridge: Cambridge University Press.

Edge, J. (2003). TEFL and International Politics: A Personal Narrative. *IATEFL Issues, 175*, 10-11.

Ellis, N. C., & Simpson, R. (2005). *An Academic Formulas List (AFL): Extraction, Validation, Prioritization.* Paper presented at the Phraseology 2005, Université Louvain-la-Neuve, Belgium, 14 October 2005.

Erman, B. (2007). Cognitive Processes as Evidence of the Idiom Principle. *International Journal of Corpus Linguistics, 12*(1), 25-53.

Erman, B., & Warren, B. (2000). The Idiom Principle and the Open Choice Principle. *Text, 20*(1), 29-62.

Eskildsen, S. W., & Cadierno, T. (2006, May 19-20, 2006). *Are Recurring Multi-Word Expressions Really Syntactic Freezes? Second Language Acquisition from*

the Perspective of Usage-Based Linguistics. Paper presented at the First Nordic Conference on Syntactic Freezes, University of Joensuu, Finland.
Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930-1955. *Studies in Linguistic Analysis*, 1-32.
Flowerdew, J. (2002). Introduction: Approaches to the Analysis of Academic Discourse in English. In J. Flowerdew (Ed.), *Academic Discourse*. Harlow: Longman.
Flowerdew, L. (2004). The Argument for Using English Specialized Corpora to Understand Academic and Professional Language. In U. Connor & T. A. Upton (Eds.), *Discourse in the Professions* (pp. 11-33). Amsterdam: John Benjamins.
Fox, G. (1987). The Case for Examples. In J. M. Sinclair (Ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing* (pp. 137-149). London: Collins ELT.
Francis, W. N., & Kučera, H. (1964/1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.* Providence, Rhode Island: Brown University Department of Linguistics.
Franz, A., & Brants, T. (2006). All Our N-Gram Are Belong to You. Retrieved 25 January 2008, from http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
Funk, C. E. (1950). *A Hog on Ice and Other Curious Expressions*. London: John Murray.
Gardner, D. (2007). Validating the Construct of Word in Applied Corpus-Based Vocabulary Research: A Critical Survey. *Applied Linguistics, 28*(2), 241-265.
Gledhill, C. (1999). The Phraseology of Rhetoric, Collocations and Discourse in Cancer Research Abstracts. In C. Barron & N. Bruce (Eds.), *Knowledge and Discourse: Proceedings of the International Multidisciplinary Conference. Hong Kong: 18-21 June 1996*. Hong Kong: University of Hong Kong.
Gledhill, C. (2000a). *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.
Gledhill, C. (2000b). The Discourse Function of Collocation in Research Article Introductions. *English for Specific Purposes, 19*, 115-135.
Goldberg, A. E. (2003). Constructions: A New Theoretical Approach to Language. *Trends in Cognitive Sciences, 7*(5), 219-224.
Greaves, C. (2005). *Introduction to ConcGram.* Paper presented at the Tuscan Word Centre Workshop, 25-29 June 2005, Certosa di Pontignano, Italy.
Groom, N. W. (2002). *A Corpus Study of the Patent Specification Genre.* Unpublished MPhil, University of Birmingham.
Groom, N. W. (2007). *Phraseology and Epistemology in Humanities Writing: A Corpus-Driven Study.* Unpublished PhD, University of Birmingham, Birmingham.
Hakuta, K. (1974). Prefabricated Patterns and the Emergence of Structure in Second Language Acquisition. *Language Learning, 24*(2), 287–297.
Halliday, M. A. K. (1961). Categories of the Theory of Grammar. *Word, 17*(3), 241-292.
Halliday, M. A. K. (1966). Lexis as a Linguistic Level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth* (pp. 148-162). London: Longmans.
Halliday, M. A. K. (1973). *Explorations in the Functions of Language*. London: Edward Arnold.
Halliday, M. A. K. (1978). *Language as Social Semiotic*. London: Edward Arnold.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2 ed.). London: Edward Arnold.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The Linguistic Sciences and Language Teaching*. London: Longmans.
Handford, M., & McCarthy, M. (2004). Invisible to Us: A Preliminary Corpus-Based Study of Spoken Business English In U. Connor & T. A. Upton (Eds.), *Discourse in the Professions: Perspective from Corpus Linguistics* (pp. 167-201). Amsterdam: John Benjamins.
Harwood, N. (2005). 'We Do Not Seem to Have a Theory . . . The Theory I Present Here Attempts to Fill This Gap': Inclusive and Exclusive Pronouns in Academic Writing. *Applied Linguistics, 26*(3), 343-375.
Harwood, N., & Hadley, G. (2004). Demistifying Institutional Practices: Critical Pragmatism and the Teaching of Academic Writing. *Journal of English for Specific Purposes, 23*, 355-377.
Hewings, M. (1993). The End! How to Conclude a Dissertation. In G. M. Blue (Ed.), *Language, Learning and Success: Studying through English* (pp. 105-112). London: MacMillan.
Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
Hoey, M. (1997). From Concordance to Text Structure: New Uses for Computer Corpora. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *Practical Applications in Language Corpora* (pp. 2-23). Łódź: Łódź University Press.
Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language.* London: Routledge.
Hopkins, A., & Dudley-Evans, A. (1988). A Genre-Based Investigation of the Discussion Sections in Articles and Dissertations. *English for Specific Purposes, 7*, 113-122.
Hopps, D. (2002). Wanderers Leave Leeds with No Defence. *The Guardian, 18th November 2002*.
Howarth, P. A. (1996). *Phraseology in English Academic Writing*. Tübingen: Max Niemeyer.
Howarth, P. A. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics, 19*(1), 24-44.
Hundt, M., Sand, A., & Rainer, S. (1998). *Manual of Information to Accompany the Freiburg - LOB Corpus of British English ('FLOB')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.
Hundt, M., Sand, A., & Skandera, P. (1999). *Manual of Information to Accompany the Freiburg - Brown Corpus of American English ('FROWN')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.
Hunston, S. (1993). Evaluation and Ideology in Scientific Writing. In M. Ghadessy (Ed.), *Register Analysis: Theory and Practice*. London: Pinter.
Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
Hunston, S. (2007). *Beyond Patterns: What Is Often Said in English.* Paper presented at the 1st Aston Corpus Symposium, Aston University, 4 May 2007.
Hunston, S., & Francis, G. (1996a). *COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.
Hunston, S., & Francis, G. (1996b). *COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.
Hunston, S., & Francis, G. (2000). *Pattern Grammar*. Amsterdam: Benjamins.

Hyland, K. (1996a). Nurturing Hedges in the ESP Curriculum. *System, 24*(4), 477-490.
Hyland, K. (1996b). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication, 13*(2), 251-281.
Hyland, K. (1998a). Boosting, Hedging and the Negotiation of Academic Knowledge. *Text, 18*(3), 349-382.
Hyland, K. (1998b). Persuasion and Context: The Pragmatics of Academic Metadiscourse. *Journal of Pragmatics, 30*, 437-455.
Hyland, K. (2000). *Disciplinary Discourses*. Harlow: Longman.
Hyland, K. (2003). Dissertation Acknowledgements: The Anatomy of a Cinderella Genre. *Written Communication, 20*(3), 242-268.
Hyland, K. (2004). Graduates' Gratitude: The Generic Structure of Dissertation Acknowledgements. *English for Specific Purposes, 23*, 303-324.
Hyland, K. (2008). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes, 27*, 4-21.
Hymes, D. H. (1971). *On Communicative Competence.* Philadelphia: University of Pennsylvania Press.
Jack, A. (2004). *Red Herrings and White Elephants*. London: Metro Books.
Jackson, H., & Zé Amvela, E. (2000). *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. London: Cassell.
Jasienski, M. (2006). It's Incredible How Often We're Surprised by Findings. *Nature, 440*, 1112.
Johansson, S., Leech, G. N., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* Oslo: University of Oslo Department of English.
Johns, T. F. (1986). MicroConcord: A Language-Learner's Research Tool. *System, 14*(2), 151-162.
Johns, T. F. (1996). Kibbitzer 12: Offering Insights. Retrieved 23 November, 2007, from http://www.eisu2.bham.ac.uk/Webmaterials/kibbitzers/kibbitzer12.htm
Jordan, R. R. (1990). *Academic Writing Course* (2nd ed.). London: Longman.
Jordan, R. R. (1999). *Academic Writing Course* (3rd ed.). London: Longman.
Kachru, B. B. (Ed.). (1992). *The Other Tongue: English across Cultures.* Urbana:: University of Illinois Press.
Kilgarriff, A. (2002). Re: [Corpora-List] ACL Proceedings Paper in the American National Corpus. from http://torvald.aksis.uib.no/corpora/2002-3/0239.html
Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Eleventh Euralex International Congress* (pp. 105-116). Lorient: Université de Bretagne-Sud.
Kjellmer, G. (1984). Some Thoughts on Collocational Distinctiveness. In J. Aarts & W. Meijs (Eds.), *Corpus Linguistics.* Amsterdam: Rodopi.
Koester, A. (2004). *The Language of Work.* London: Routledge.
Kress, G. (1989). Texture and Meaning. In R. Andrews (Ed.), *Narrative and Argument* (pp. 9-21). Milton Keynes: Open University Press.
Krishnamurthy, R. (1987). The Process of Compilation. In J. M. Sinclair (Ed.), *Looking Up* (pp. 62-86). London: Collins ELT.
Kuiper, K. (2004). Formulaic Performance in Conventionalised Varieties of Speech. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 37-54). Amsterdam: John Benjamins.
Lee, D. (2001). Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology, 5*(3), 37-72.

Mackenzie, I. (2000). Improvisation, Creativity, and Formulaic Language. *Journal of Aesthetics and Art Criticism., 58*(2), 173-179.
Maconie, S. (2007). *Pies and Prejudice: In Search of the North*. London: Ebury Press.
Malinowski, B. (1923). The Problem of Meaning in Primitive Languages. In C. K. Ogden & I. A. Richards (Eds.), *Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism* (pp. 451-510). London: Kegan Paul, Trench, Trubner & Company.
Martin, J. R. (2001). Language, Register and Genre. In A. Burns & C. Coffin (Eds.), *Analysing English in a Global Context: A Reader*. London: Routledge.
Mason, O. (1997). The Weight of Words: An Investigation of Lexical Gravity. In *Proceedings of PALC '97* (pp. 361-375). Łódź: Łódź University Press.
Mauranen, A. (2004). *Formulaic Sequences in Lingua Franca English.* Paper presented at the 37th BAAL Conference, 10 September 2004, Kings College, London.
McCarthy, M. (2001). *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.
McCarthy, M., & O'Keeffe, A. (2003). What's in a Name? Vocatives in Casual Conversations and Radio Phone-in Calls. In C. Meyer & P. Leistyna (Eds.), *Corpus Analysis: Language Structure and Language Use* (Vol. 153-185). Rodopi: Amsterdam.
McIntosh, A. (1961). Patterns and Ranges. *Language, 37*(3), 325-337.
Moon, R. (1987). The Analysis of Meaning. In J. M. Sinclair (Ed.), *Looking Up* (pp. 86-103). London: Collins.
Moon, R. (1994). *Fixed Expressions and Text: A Study of the Distribution and Textual Behaviour of Fixed Expressions in English.* Unpublished PhD, University of Birmingham, UK.
Moon, R. (1996). Introduction. In R. Moon (Ed.), *Collins COBUILD Dictionary of Idioms.* London: HarperCollins.
Moon, R. (1997). Vocabulary Connections: Multi-Word Items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.
Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus Based Approach*. Oxford: Clarendon Press.
Moon, R. (2001). The Distribution of Idioms in English. *SILTA, (Studi Italiani di Linguistica Teorica e Applicata)*(2), 229-241.
Moon, R. (2007). Sinclair, Lexicography, and the COBUILD Project: The Application of Theory. *International Journal of Corpus Linguistics, 12*(2), 159-181.
Moore, T., & Carling, C. (1988). *The Limitations of Language*. London: Macmillan.
Murison-Bowie, S. (1993). *MicroConcord Manual*. Oxford: Oxford University Press.
Myers, G. (1990). *Writing Biology: Texts in the Social Construction of Scientific Knowledge*. Madison: University of Wisconsin Press.
Nattinger, J. R., & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.
Nesselhauf, N. (2003). What Are Collocations? In D. Allerton, N. Nesselhauf & P. Skandera (Eds.), *Phraseological Units. Basic Concepts and Their Application* (pp. 1-22). Basel: Schwabe.
Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.
O'Brien, F. (1968). *The Best of Myles*. London: Picador.

O'Donnell, M. B. (2007). *The Structure of Newspaper Articles: Beginnings, Middles and Ends.* Paper presented at the 1st CLARET Workshop, 30 November 2007, University of Liverpool.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom.* Cambridge: Cambridge University Press.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Oakey, D. J. (1998). *A Study of Some Lexical Phrases in a Corpus of Academic English.* Unpublished MEd Dissertation, University of Leeds, Leeds.

Oakey, D. J. (2002a). Formulaic Language in English Academic Writing: A Corpus-Based Study of the Formal and Functional Variation of a Lexical Phrase in Different Academic Disciplines in English. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 111-130). Amsterdam: John Benjamins.

Oakey, D. J. (2002b). Lexical Phrases for Teaching Academic Writing in English: Corpus Evidence. In S. Nuccorini (Ed.), *Phrases and Phraseology - Data and Descriptions* (pp. 85-105). Bern: Peter Lang.

Oakey, D. J. (2005). Academic Vocabulary in Academic Discourse: The Phraseological Behaviour of EVALUATION in Economics Research Articles. In E. Tognini-Bonelli & G. Del Lungo Camiciotti (Eds.), *Strategies in Academic Discourse* (pp. 169-183). Amsterdam: John Benjamins.

Orwell, G. (1946). Politics and the English Language. *Horizon,* April, 1946.

Parris, M. (2007). *Not My Words, Mr Speaker.* London: BBC Radio 4. Retrieved 25 January 2008, from http://news.bbc.co.uk/1/hi/programmes/the_westminster_hour/6991605.stm.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Amsterdam: John Benjamins.

Pawley, A., & Syder, F. H. (1983). Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication.* (pp. 191-226). London: Longman.

Peacock, M. (2007). *A Comparative Study of Introductory 'It' in Research Articles across Eight Disciplines.* Paper presented at the Corpus Linguistics 2007 Conference, 28 July 2007, University of Birmingham.

Pennycook, A. (1994). *The Cultural Politics of English as an International Language.* London: Longman.

Peters, A. (1983). *The Units of Language Acquisition.* Cambridge: Cambridge University Press.

Phillipson, R. (1988). *Linguistic Imperialism.* Oxford: Oxford University Press.

Piao, S., Rayson, P., Archer, D., Wilson, A., & McEnery, T. (2003). *Extracting Multiword Expressions with a Semantic Tagger.* Paper presented at the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan.

Prodromou, L. (2006). Defining the 'Successful Bilingual Speaker' of English. In R. Rubdy & M. Saraceni (Eds.), *English in the World: Global Rules, Global Roles* (pp. 51-70). London: Continuum.

Quirk, R., & Greenbaum, S. (1973). *A University Grammar of English.* Harlow: Longman.

Rayson, P. (2008). Wmatrix: A Web-Based Corpus Processing Environment. Lancaster: Computing Department, Lancaster University.

Robins, R. H. (1961). John Rupert Firth. *Language, 37*(2), 191-199.

Schiffrin, D. (1993). *Approaches to Discourse*. Oxford: Blackwell.
Schmitt, N. (Ed.). (2004). *Formulaic Sequences*. Amsterdam: John Benjamins.
Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are Corpus-Derived Recurrent Clusters Psycholinguistically Valid? In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127-151). Amsterdam: John Benjamins.
Scott, M. (1999). Wordsmith Tools (Version 3). Oxford: Oxford University Press.
Scott, M. (2005). Wordsmith Tools (Version 4). Oxford: Oxford University Press.
Scott, M., & Johns, T. F. (1993). MicroConcord. Oxford: Oxford University Press.
Simpson, J., & Weiner, E. (Eds.). (1989). *Oxford English Dictionary* (2 ed.). Oxford: Oxford University Press.
Sinclair, J. M. (1966). Beginning the Study of Lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth*. London: Longman.
Sinclair, J. M. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.
Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Sinclair, J. M. (1996). The Search for Units of Meaning. *Textus, 9*, 75-106.
Sinclair, J. M. (1998). The Lexical Item. In E. Weigand (Ed.), *Contrastive Lexical Semantics*. Amsterdam: John Benjamins.
Sinclair, J. M. (2001). Preface. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT*. Amsterdam: John Benjamins.
Sinclair, J. M. (2005). Corpus and Text — Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books.
Sinclair, J. M. (Ed.). (2004). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
Sinclair, J. M. (Ed.). (2006). *Collins COBUILD Advanced Learner's English Dictionary 5th Edition*. Glasgow: HarperCollins.
Sinclair, J. M., & Coulthard, M. (1975). *Towards an Analysis of Discourse*. Oxford: Oxford University Press.
Smith, Z. (2007). Fail Better. *The Guardian, January 13 2007*
Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Oxford: Basil Blackwell.
Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Swales, J. M. (1981). *Aspects of Article Introductions* (Aston Research Report No. 1). Birmingham: University of Aston.
Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.
Swales, J. M. (2001). EAP-Related Linguistic Research: An Intellectual History. In J. Flowerdew & M. Peacock (Eds.), *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.
Swales, J. M. (2002). Integrated and Fragmented Worlds: EAP Materials and Corpus Linguistics. In J. Flowerdew (Ed.), *Academic Discourse* (pp. 150-164). Harlow: Longman.
Swan, M. (1995). *Practical English Usage* (2nd ed.). Oxford: Oxford University Press.
Thurstun, J., & Candlin, C. (1997). *Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: National Centre for English Language Teaching and Research.

Timmis, I. (2002). Native-Speaker Norms and International English: A Classroom View. *English Language Teaching Journal, 56*(3), 240-249.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The Eyes Have It: An Eye-Movement Study into the Processing of Formulaic Sequences. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 153-172). Amsterdam: John Benjamins.

Unicode. (2008). Ligatures, Digraphs and Presentation Forms. Retrieved 3 January, 2008, from http://unicode.org/faq/ligature_digraph.html#5

Van Lancker, D. (1974). Heterogeneity in Language and Speech. *UCLA Working Papers in Phonetics, 29*.

West, G. K. (1980). That-Nominal Constructions in Traditional Rhetorical Divisions of Scientific Research Papers. *TESOL Quarterly, 14*, 483-489.

Willis, D. (1995). Review of Lexical Phrases and Language Teaching. *English Language Teaching Journal, 49*(1), 87-90.

Wray, A. (1999). Formulaic Language in Learners and Native Speakers. *Language Teaching, 32*, 213-231.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The Functions of Formulaic Language: An Integrated Model. *Language and Communication, 20*, 1-28.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, Mass.: MIT Press.

Corpora referred to in this study

Academic Prose Corpus (Hyland 2008)

Bank of English (2008)
The Bank of English corpus is jointly owned by HarperCollins Publishers and the University of Birmingham. In 2008 the corpus stands at 524 million words; Figures quoted in this thesis are from the 450 million word 2001 version.

Brigham Young University Corpus of American English (Davies 2008) Retrieved 28 February 2008, from http://www.americancorpus.org/

British National Corpus World Edition (BNC) (2000) Oxford: Oxford University Press

Cambridge International Corpus (2007) Cambridge: Cambridge University Press Retrieved 1 January 2008, from http://www.cambridge.org/elt/corpus/international_corpus.htm

Cancer Research Article Corpus (Gledhill 2000)

Collins Word Web (2007) Retrieved 1 January 2008, from http://www.collins.co.uk/books.aspx?group=180

FLOB Corpus (Hundt et al. 1998)

FROWN Corpus (Hundt et al. 1999)

Google Public Web Training Corpus (Franz & Brants 2006)

Guardian News Archive (Scott 2007)

Humanities Corpus (Groom 2007)

London Oslo Bergen Corpus (LOB) (Johansson et al. 1978)

Longman Corpus of Spoken and Written English (Biber et al. 1999)

MicroConcord B Corpus (1993) Oxford: Oxford University Press

Research Article Corpus (Hyland 1998)

Standard Corpus of Present Day American English (Brown) (Francis & Kučera 1964/1979)

T2K-SWAL Corpus (Biber, Conrad, Reppen et al. 2004)

Vienna-Oxford International Corpus of English (VOICE) (Seidlhofer 2001) Retrieved 1 January 2008, from http://www.univie.ac.at/voice/ Seidlhofer, B. (2001).

Appendices

Appendix 1 Letter to university departments and their replies

Appendix 1.1: Initial letter

From: David J. Oakey

To: Professor X

Subject: English in [your academic subject] journals

Dear Professor X,

I am a lecturer in the English for International Students Unit, providing academic English support to non-native speaking international students. I am conducting some research into the English of [your academic subject]. I hope to look at some journals in the field, and I would like to ask your opinion about the best ones to look at. I am hoping to focus on those which are most prestigious, widely read, but not too specialised in a particular sub-discipline. I have a list of journals below used by another researcher but I'm not sure whether they are the same ones used at Birmingham:

[list of journals from Hyland Journal Corpus]

I'd be grateful if you could help at all, or could pass this email on to one of the academic staff whom you think might be able to help, or let me know who else in the department to contact.

Yours sincerely,

David Oakey

Appendix 1.2 Responses from departments: Cell Biology

The most prestigous Cell Biology journals are;

xCell (not just cell biology but some)
yEMBO journal (as above)
yNature Cell Biology
yJournal of Cell Biology
yJournal of Cell Science
Current Biology (this is general journal but with lots of Cell Bio in)
Molecular and Cellular Biology
yMolecular Biology of the Cell
Trends in Cell Biology (for reviews).

Appendix 1.3 Responses from departments: Electrical and Electronic Engineering

Attached is a pdf file containing a list of journals. This is compiled by the survey company 'Evidence' based on the last research assessment exercise. The report is from research and business enterprises.

UoA 29 Electrical and Electronic Engineering

Top 20 journals submitted to this UoA - by frequency of occurrence in RA2

ISSN	Journal title	Rank on papers - 5 and 5*	Rank on papers - all units	Rank on average impact	Papers - 5 and 5*	Papers - all units	Average impact
0013-5194	Electronics Letters	1	1	161	142	208	2.10
0003-6951	Applied Physics Letters	2	3	14	67	74	7.19
0021-8979	Journal Of Applied Physics	3	2	59	62	76	3.83
0018-9480	Ieee Transactions On Microwave Theory And	4	5	143	58	68	2.32
1350-2352	Iee Proceedings-electric Power Applications	5	6	337	54	62	0.84

Appendix 1.4 Responses from departments: Mechanical Engineering

I am replying your recent enquiry regarding the above.

From your list these two are available in electronic format via Talis:

'Mechanics & Machine Theory' and 'International Journal of Mechanical Sciences' . They are published by Elsevier who is regarded very highly in the profession.

The following three titles are parts of the 'Proceedings of the Institute of Mechanical Engineers' which is regarded as almost a bible for Mechanical Engineers and very prestigious. They are available in hardbound format.

'Journal of Process Mechanical Engineering' Part E q TJ1.14
'Journal of Mechanical Engineering Science' Part C q TJ1.14
'Journal of Engineering Manufacture' Part B q TJ1.14

The other two titles are not available in the University Library but I think there is more than enough choice for the students from the above.

Please contact me if I can help further.

Appendix 1.5 Responses from departments: Nuclear Physics

The journals given are all respected Journals. Let me give a little more information.

Physical Review Letters is the most prestigious, but carries articles on all aspects of physics not just nuclear physics. It is also an american journal as is Physical Review C, Nuclear Physics.

Journal of Physics G: Nuclear and Particle Physics is a British journal which has become more significant in recent years.

The other two are North Holland publications.
Nuclear Instruments and Methods in Physics Research, is an odd man out as it

deals with instrumentation rather than physics per se.

Another North Holland journal is Nuclear Physics A. This was the preferred journal for UK publications but is being eclipsed somewhat by J. Phys G. Phys Rev C and Phys Rev Lett have always been popular places to publish.

Both the american journals charge authors to puplish (although one can get round these costs) and by doing so keep their costs to readers down.
The others journals do not have page charges.

Appendix 1.6 Responses from departments: Economics

I wish I could share with you a paper I recently refereed and my referee report, but that is confidential. It was a massive statistical analysis of language in six journals, marred, in my view, by a peculiar choice of journals.

It is difficult to answer your question simply. At one time, the EJ was probably the major journal, but it has declined. There are numerous studies of journal rankings in economics - and you could look at some of those rankings to get an idea of prestige.

When people think of the top "general" general journals, the ones that they usually think of are the American Economic Review, the Quarterly Journal of Economics, and the Journal of Political Economy. (Economics is dominated by the US, even more than many other disciplines, so they are all American). Econometrica is also usually considered to be a top journal, but that is more specialised on mathematical and quantitative work (not that you will think the other three non-mathematical!).

In the rankings the EJ is usually, I believe, around 10 or so, though its editors have been actively trying to remedy this situation.

The AER (above) is the one with refereed submitted articles. The JEL(iterature) is also produced by the American Economic Association, and is their abstracting journal, containing a small number of articles that are of a survey nature (broadly

interpreted). The JEP(erspectives) is also an AEA journal, but deliberately aimed to present economic ideas in a way that is accessible to students, for whom the main technical journals are too difficult. The EJ, combines this type of material in the main journal, though recently it began to put this type of material into separately-paginated issues so as not to "dilute" the main refereed articles.

After these, and the Brookings Papers, a policy journal, which are at the top of most rankings, the field journals come next - covering various applied fields. There are other general journals, but a lot of these come lower down and are considered places that most economists would publish in if they cannot get into the top journals (or if they are simply not maximizing or are publishing material that is of only local (ie non US) interest. Of these, the European Economic Review is a good one (sponsored by the European Economic Association). The EJ is up at the top of this category, if it is not higher. The Southern Economic Journal also often gets ranked highly (another US one) Other regional US journals come a bit lower (Economic Inquiry, Eastern (quite low that), Atlantic (ditto)).

JET(heory) is in a sense a specialist journal, but it is mathematical theory that defines the specialism, so there is a sense in which it is general. Economic Theory is probably similar, though I don't know it so well.

Agricultural Economics is really a different discipline - historically that has always been the case. In some places Ag econ is in economics departments, but it has historically been pretty separate. Most economists would not read it or even know anyone in that field.

Journal of Socio-Economics: I am not sure about this one, but I suspect it is either a sort of "fringe" journal, or even one that might be thought of as Sociology rather than economics. It certainly is not either a central journal or a highly ranked one.

As you will tell, economics is in a sense a very hierarchical discipline when it comes to journals. I once did a survey of the literature on journal rankings and other exercises, though did not publish it. I could dig it out, or else you could probably find some examples of the genre looking through the Journal of Economic Perspectives.

My advice on journal selection would be that it would depend on the questions you want to ask. The problem with the paper I mentioned at the start was that it had some wonderful techniques and must have analyzed a corpus of millions of words, but it ended up proving what anyone who knew the literature would have predicted with 15 minutes careful thought. The questions are very important. I assume you are aware that the work with Dudley-Evans and Henderson was the tip of an iceberg of work on language in economics. If not, you should consult Tony Dudley-Evans (in English, who did not retire that long ago) as much of it was done in Birmingham - they had several research projects that produced several reports and articles. Henderson is still around as Head of continuing education (if that is what it is still called). Tim Johns was also involved in some of it.

For your purposes, the selection of journals may be less critical. The author I mentioned (and I admit that I had to say that I did not think it suitable for the journal to which it had been sent - I did not think it would hold the interest of economists given its length in relation to what it said) wanted to make comparisons within the subject, so his sample was much more critical.

If you wanted High-prestige refereed articles, the QJE, JPE, AER and EJ (you might want to select certain parts of the EJ, depending on your purposes - it contains book reviews and some JEP type material as well as the refereed articles) would be fine. The JEP is also central if you bear in mind its aim at a less technical audience including students. The JEL is the standard bibliographic one. Brookings Papers on Economic Activity would be a good policy-related choice. Oxford Review of Economic Policy would be a comparable UK one, though less prestigious I think.

Appendix 1.7 Responses from departments: Marketing

What an interesting project. All the journals you have below are prestigious and well respected. I have ranked these (number next to Journal below). It is a subjective assessment as we do not have an official journal ranking. Incidentally I would add

European Journal of Marketing which is slightly more prestigious than Journal of Marketing Management. Regards Leslie

002 Journal of Marketing 1y
003 Journal of Marketing Research 2y
004 Journal of the Academy of Marketing Science 3y
Marketing Science 4y
International Journal of Research in Marketing 5n
European Journal of Marketing
Journal of Marketing Management 6y
001 Journal of Marketing Communication 7y

Appendix 1.8 Responses from departments: Philosophy

These journals are the most prestigious.
You could add to your list: The philosophical review

Appendix 1.9 Responses from departments: Sociology

These journals are all sensible ones to use and could be retained if you wished to retain comparability with the previous study.

Otherwise, I would recommend AJS and Social Problems as your two American Journals (both highly prestigious, but the first with a professional agenda and the second with a remit to engage with public issues) and the British Journal of Sociology (as broadly comparable with the AJS) and Sociology (rather than Sociological Review) as the top-ranked British journal.

if it is a question of rankings then, American Journal of Sociology, American Sociological review and social forces are the US top three, while British Journal of Sociology and Sociology are the top British two.

Appendix 2: Summary of the isolexical and isotextual versions of the corpus

Discipline	Journal	Isolexical version		Isotextual version	
		Texts	Tokens	Texts	Tokens
Electrical Engineering	*Electronics Letters*	760	1,153,020	100	152,484
	Journal of Applied Physics	225	1,215,514	100	542,619
	IEEE Transactions on Microwave Theory and Techniques	213	1,171,677	100	570,175
	Applied Physics Letters	466	1,191,938	100	255,320
	IEE Proceedings -- Electric Power Applications	245	1,251,805	100	529,799
Nuclear Physics	*Journal of Physics G: Nuclear and Particle Physics*	259	1,221,472	100	553,513
	Physical Review Letters	353	1,151,874	100	334,611
	Physical Review C, Nuclear Physics	184	1,287,367	100	722,140
	Nuclear Physics A	143	1,137,837	100	823,352
	Nuclear Physics B	146	1,201,138	100	791,450
Mechanical Engineering	*Journal of Fluids Engineering*	171	1,067,269	100	636,689
	Mechanism and Machine Theory	218	1,094,820	100	521,763
	Journal of Engineering Manufacture	205	1,080,586	100	527,780
	International Journal of Mechanical Sciences	196	1,057,634	100	539,653
	Journal of Biomechanics	178	1,045,090	100	572,317
Cell Biology	*Journal of Cell Biology*	114	1,070,665	100	936,473
	Journal of Cell Science	121	1,045,245	100	868,431
	Nature Cell Biology	177	1,079,429	100	611,358
	Molecular Biology of the Cell	111	1,037,615	100	934,823
	European Molecular Biology Organization Journal	118	1,015,941	100	859,748
Economics	*Journal of Economic Theory*	109	1,055,025	100	972,254
	Economic Journal	119	1,055,996	100	913,054
	American Economic Review	97	1,055,245	100	1,101,638
	Quarterly Journal of Economics	70	1,068,130	100	1,515,275
	Journal of Political Economy	82	1,026,680	100	1,189,145
Philosophy	*Erkenntnis*	117	1,037,525	100	899,020
	Inquiry	114	1,047,527	100	938,382
	Philosophy	129	1,031,376	100	812,087
	The Philosophical Review	72	1,024,927	100	1,433,013
	The Philosophical Quarterly	135	1,032,606	100	758,223
Marketing	*Journal of Marketing*	81	1,022,229	100	1,266,016
	Journal of Marketing Research	96	1,053,618	100	1,094,675
	Journal of the Academy of Marketing Science	109	1,061,118	100	959,213
	Marketing science	92	1,065,549	100	1,153,401
	European Journal of Marketing	134	1,072,050	100	802,097
Sociology	*American Journal of Sociology*	66	1,059,862	100	1,620,002
	The Sociological Review	107	1,025,828	100	960,777
	Current Sociology	132	1,025,750	100	767,649
	British Journal of Sociology	115	1,037,081	100	914,893
	Annual Review of Sociology	85	1,032,764	100	1,189,923
Totals		**6664**	**43,468,822**	**4000**	**33,352,770**

Appendix 3: Detailed information on the composition of the isolexical and isotextual versions of the research article corpus

Tokens are counted using WordSmith Tools 4 with the following conditions: numbers included; hyphens break words. For the isotextual version the column "details" lists the articles included from each volume.

Appendix 3.1 Cell biology

			Isolexical version			Isotextual version			
Journal	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Journal of Cell Biology	2003	162	17	163,798	html	3	709-711	29,690	html
	2003	163	97	906,867	html	97		906,783	html
	Totals		**114**	**1,070,665**		**100**		**936,473**	
Journal of Cell Science	2003	116	121	1,045,245	html	100	1013-15-17	868,431	html
	Totals		**121**	**1,045,245**		**100**		**868,431**	
Nature Cell Biology	2001	3	100	615,277	pdf	0		0	n/a
	2002	4	77	464,152	pdf	77		464,152	pdf
	2003	5	0	0	pdf	23		147,206	pdf
	Totals		**177**	**1,079,429**		**100**		**611,358**	
Molecular Biology of the Cell	2003	14	111	1,037,615	html	100	1001-1234	934,823	html
	Totals		**111**	**1,037,615**		**100**		**934,823**	
European Molecular Biology Organization Journal	2003	22	118	1,015,941	html	100	1919-2219	859,748	html
	Totals		**118**	**1,015,941**		**100**		**859,748**	
	Full Totals		**641**	**5,248,895**		**500**		**4,210,833**	

Appendix 3.2 Electrical Engineering

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Electronics Letters	2003	39	391	590,761	pdf	0		0	n/a
	2004	40	369	562,259	pdf	100	1-303	152,484	pdf
	Totals		**760**	**1,153,020**		**100**		**152,484**	
Journal of Applied Physics	2003	94	225	1,215,514	html	100	1 only	542,619	html
	Totals		**225**	**1,215,514**		**100**		**542,619**	
IEEE Transactions on Microwave Theory and Techniques	2003	51	35	182,721	pdf	0		0	n/a
	2004	52	178	988,956	pdf	100	1-320	570,175	pdf
	Totals		**213**	**1,171,677**		**100**		**570,175**	
Applied Physics Letters	2003	84	187	478,478	html	0		0	n/a
	2004	85	279	713,460	html	100	1-247	255,320	html
	Totals		**466**	**1,191,938**		**100**		**255,320**	
IEE Proceedings -- Electric Power Applications	2001	148	28	134,838	pdf	0		0	n/a
	2002	149	49	232,363	pdf	0		0	n/a
	2003	150	90	467,019	pdf	22	512-617	112,246	pdf
	2004	151	78	417,585	pdf	78	1 to 5	417,553	pdf
	Totals		**245**	**1,251,805**		**100**		**529,799**	
	Full Totals		**1,909**	**5,983,954**		**500**		**2,050,397**	

Appendix 3.3 Mechanical Engineering

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Journal of Fluids Engineering	2001	123	57	328,843	html	0		0	n/a
	2002	124	114	738,426	html	100	118-426	636,689	html
	Totals		**171**	**1,067,269**		**100**		**636,689**	
Mechanism and Machine Theory	2001	36	14	57,430	html	0		0	n/a
	2002	37	83	406,347	html	0		0	n/a
	2003	38	57	296,337	html	36	405-1211	187,057	html
	2004	39	64	334,706	html	64		334,706	html
	Totals		**218**	**1,094,820**		**100**		**521,763**	
Journal of Engineering Manufacture	2002	216	85	438,320	pdf	0		0	n/a
	2003	217	120	642,266	pdf	100	211-1213	527,780	pdf
	Totals		**205**	**1,080,586**		**100**		**527,780**	
International Journal of Mechanical Sciences	2002	44	40	203,454	html	0		0	n/a
	2003	45	105	583,548	html	49	610-1207	269,021	html
	2004	46	51	270,632	html	51		270,632	html
	Totals		**196**	**1,057,634**		**100**		**539,653**	
Journal of Biomechanics	2003	36	44	284,129	html	0		0	n/a
	2004	37	134	760,961	html	100	311-915	572,317	html
	Totals		**178**	**1,045,090**		**100**		**572,317**	
	Full Totals		**968**	**5,345,399**		**500**		**2,798,202**	

Appendix 3.4 Nuclear physics

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Journal of Physics G: Nuclear and Particle Physics									
	1998	24	110	471,779	pdf	0		0	n/a
	1999	25	149	749,693	pdf	100	301-1206	553,513	pdf
	Totals		**259**	**1,221,472**		**100**		**553,513**	
Physical Review Letters	1999	82	83	263,944	pdf	0		0	n/a
	1999	83	78	250,297	pdf	0		0	n/a
	2000	84	60	198,098	pdf	0		0	n/a
	2000	85	66	216,311	pdf	34	1304-2603	111,387	n/a
	2001	86	66	223,224	pdf	66		223,224	na
	Totals		**353**	**1,151,874**		**100**		**334,611**	
Physical Review C, Nuclear Physics		57	184	1,287,367	pdf	100	533-663	722,140	pdf
	Totals		**184**	**1,287,367**		**100**		**722,140**	
Nuclear Physics A	2003	724	28	201,276	pdf	0		0	n/a
	2003	725	16	120,743	pdf	1	116	7,534	pdf
	2003	726	17	138,202	pdf	17		138,202	pdf
	2003	727	20	187,136	pdf	20		187,136	pdf
	2003	728	25	185,101	pdf	25		185,101	pdf
	2003	729	10	85,027	pdf	10		85,027	pdf
	2004	730	27	220,352	pdf	27		220,352	pdf
	Totals		**143**	**1,137,837**		**100**		**823,352**	
Nuclear Physics B	2003	648	23	191,888	html	0		0	n/a
	2003	649	21	179,376	html	0		0	n/a
	2004	691	12	94,530	html	10	103-303	55,958	html
	2004	692	15	123,431	html	15		123,431	html
	2004	693	11	84,654	html	11		84,654	html
	2004	694	16	134,397	html	16		134,397	html
	2004	695	14	78,116	html	14		78,138	html
	2004	696	17	147,836	html	17		147,913	html
	2004	697	17	166,910	html	17		166,959	html
	Totals		**146**	**1,201,138**		**100**		**791,450**	
	Full totals		**1,085**	**5,999,688**		**500**		**3,225,066**	

Appendix 3.5 Economics

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Journal of Economic Theory	2000	91	13	113,045	pdf	4	204-207	30,274	pdf
	2000	92	12	134,547	pdf	12		134,547	pdf
	2000	93	13	109,602	pdf	13		109,602	pdf
	2000	94	11	103,146	pdf	11		103,146	pdf
	2000	95	10	94,492	pdf	10		94,492	pdf
	2001	96	11	98,357	pdf	11		98,357	pdf
	2001	97	13	129,431	pdf	13		129,431	pdf
	2001	98	15	137,021	pdf	15		137,021	pdf
	2001	99	11	135,384	pdf	11		135,384	pdf
	Totals		**109**	**1,055,025**		**100**		**972,254**	
Economic Journal	1999	109	59	499,898	html	40	409-809	356,956	html
	2000	110	60	556,098	html	60		556,098	html
	Totals		**119**	**1,055,996**		**100**		**913,054**	
American Economic Review	1999	89	10	96,271	tif	13	405-417	142,664	tif
	2000	90	87	958,974	tif	87		958,974	tif
	Totals		**97**	**1,055,245**		**100**		**1,101,638**	
Quarterly Journal of Economics	2002	117	30	451,960	html	40	101-410	600,538	html
	2003	118	40	616,170	html/pdf	40		616,290	html/pdf
	2004	119	0	0	n/a	20	101-210	298,447	html
	Totals		**70**	**1,068,130**		**100**		**1,515,275**	
Journal of Political Economy	2002	110	40	472,250	html	40		472,250	html
	2003	111	42	554,430	html	42		554,626	html
	2004	112	0	0	n/a	18	101-211	199,340	pdf
	Totals		**82**	**1,026,680**		**100**		**1,226,216**	
	Full Totals		**477**	**5,261,076**		**500**		**5,728,437**	

Appendix 3.6 Marketing

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Journal of Marketing	1999	63	0	0	n/a	19	1-405	243,787	html
	2000	64	27	325,326	html	27		325,326	html
	2001	65	25	295,388	html	25		295,388	html
	2002	66	29	401,515	html	29		401,515	html
	Totals		**81**	**1,022,229**		**100**		**1,266,016**	
Journal of Marketing Research	2000	37	19	210,335	html	23	207 to 410	251,392	html
	2001	38	40	428,961	html	40		428,961	html
	2002	39	37	414,322	html	37		414,322	html
	Totals		**96**	**1,053,618**		**100**		**1,094,675**	
Journal of the Academy of Marketing Science	1999	27	18	186,213	html	9	305-406	84,308	html
	2000	28	36	314,577	html	36		314,577	html
	2001	29	23	252,937	html	23		252,937	html
	2002	30	32	307,391	html	32		307,391	html
	Totals		**109**	**1,061,118**		**100**		**959,213**	
Marketing Science	1998	17	0	0	n/a	8	307-406	87,852	html
	1999	18	23	272,475	html	23		272,475	html
	2000	19	23	270,995	html	23		270,995	html
	2001	20	21	218,649	pdf	21		218,649	pdf
	2002	21	25	303,430	pdf	25		303,430	pdf
	Totals		**92**	**1,065,549**		**100**		**1,153,401**	
European Journal of Marketing	2000	34	27	218,284	html	0		0	**n/a**
	2001	35	64	498,552	html	57	108-1111	446,883	html
	2002	36	43	355,214	html	43		355,214	html
	Totals		**134**	**1,072,050**		**100**		**802,097**	
	Full Totals		**512**	**5,274,564**		**500**		**5,275,402**	

Appendix 3.7 Philosophy

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
Erkenntnis	2000	52	16	129,459	pdf	0		0	n/a
	2000	53	16	165,887	pdf	15		156,841	pdf
	2001	54	15	121,997	pdf	15		121,997	pdf
	2001	55	17	148,398	pdf	17		148,398	pdf
	2002	56	17	150,824	pdf	17		150,824	pdf
	2002	57	19	177,015	pdf	19		177,015	pdf
	2003	58	17	143,945	pdf	17		143,945	pdf
	Totals		**117**	**1,037,525**		**100**		**899,020**	
Inquiry	1998	41	17	136,123	pdf	3	404-406	26,978	pdf
	1999	42	24	241,723	pdf	24		241,723	pdf
	2000	43	26	230,737	pdf	26		230,737	pdf
	2001	44	22	244,015	pdf	22		244,015	pdf
	2002	45	25	194,929	pdf	25		194,929	pdf
	Totals		**114**	**1,047,527**		**100**		**938,382**	
Philosophy	1998	73	8	39,569	pdf	0		0	n/a
	1999	74	24	201,901	pdf	3	404-406	22,181	pdf
	2000	75	24	213,925	pdf	24		213,925	pdf
	2001	76	26	202,472	pdf	26		202,472	pdf
	2002	77	24	214,194	pdf	24		214,194	pdf
	2003	78	23	159,315	pdf	23		159,315	pdf
	Totals		**129**	**1,031,376**		**100**		**812,087**	
The Philosophical Review	1994	103	15	206,082	tif	15		205,012	tif
	1995	104	11	127,292	tif	11		127,292	tif
	1996	105	0	0	n/a	10		136,340	pdf
	1997	106	12	177,745	pdf	12		177,745	pdf
	1998	107	11	175,799	pdf	11		175,799	pdf
	1999	108	11	166,869	pdf	11		166,869	pdf
	2000	109	12	171,140	pdf	12		171,140	pdf
	2001	110	0	0	n/a	11		166,389	tif
	2002	111	0	0	n/a	7	101-302	106,427	tif
	Totals		**72**	**1,024,927**		**100**		**1,433,013**	
The Philosophical Quarterly	1991	41	18	142,443	tif	0		0	n/a
	1992	42	18	136,336	tif	0		0	n/a
	1993	43	17	123,080	tif	2	405-406	13,250	tif
	1994	44	16	124,269	tif	16		124,269	tif
	1995	45	19	124,883	tif	19		124,883	tif
	1996	46	0	0	n/a	16		114,226	pdf
	1997	47	17	142,604	pdf	17		142,604	pdf
	1998	48	17	121,699	pdf	17		121,699	pdf
	1999	54	13	117,292	pdf	13		117,292	pdf
	Totals		**135**	**1,032,606**		**100**		**758,223**	
	Full Totals		**567**	**5,173,961**		**500**		**4,840,725**	

Appendix 3.8: Sociology

			Isolexical			Isotextual			
	Year	Vol.	Texts	Tokens	Format	Texts	Details	Tokens	Format
American Journal of Sociology	1998	104	0	0	n/a	10	503-606	158,706	
	1999	105	38	581,495	html	38		581,495	html
	2000	106	0	0	na	24	301-606	401,434	
	2001	107	28	478,367	html	28		478,367	html
	Totals		**66**	**1,059,862**		**100**		**1,620,002**	
The Sociological Review	1999	47	28	281,342	html	21	201-407	216,291	html
	2000	48	30	282,030	html	30		282,030	html
	2001	49	21	196,172	html	21		196,172	html
	2002	50	21	192,205	html	21		192,205	html
	2003	51	7	74,079	pdf	7		74,079	pdf
	Totals		**107**	**1,025,828**		**100**		**960,777**	
Current Sociology	1999	47	6	44,814	pdf	0		0	n/a
	2000	48	21	164,798	pdf	0		0	n/a
	2001	49	38	311,023	pdf	33	106-606	262,534	pdf
	2002	50	41	304,650	pdf	41		304,650	pdf
	2003	51	26	200,465	pdf	26		200,465	pdf
	Totals		**132**	**1,025,750**		**100**		**767,649**	
British Journal of Sociology	1999	50	12	98,599	html	0		0	n/a
	2000	51	27	267,475	html	24	104-308	243,886	html
	2001	52	33	304,493	html	33		304,493	html
	2002	53	35	300,040	html	35		300,040	html
	2003	54	8	66,474	html	8		66,474	html
	Totals		**115**	**1,037,081**		**100**		**914,893**	
Annual Review of Sociology	2000	26	22	242,417	pdf	22		242,417	pdf
	2001	27	19	243,147	pdf	19		243,147	pdf
	2002	28	20	257,765	pdf	20		257,765	pdf
	2003	29	24	289,435	pdf	24		289,435	pdf
	2004	30	0	0	n/a	15		157,159	pdf
	Totals		**85**	**1,032,764**		**100**		**1,189,923**	
	Full Totals		**505**	**5,181,285**		**500**		**5,453,244**	

Appendix 4: Comparison of top 50 3-word patterns: Cell Bio.

3-word patterns in Cell Biology	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
the presence of	3,176	1	2564	1	0
in the presence	2,353	2	1939	2	0
is required for	1,708	3	1289	4	-1
the absence of	1,703	4	1364	3	1
in the absence	1,576	5	1274	5	0
the plasma membrane	1,359	6	1134	6	0
as well as	1,279	7	1018	7	0
at 37 c	1,217	8	998	8	0
in response to	1,194	9	959	9	0
at 4 c	1,106	10	891	10	0
the formation of	961	11	717	12	-1
as described previously	908	12	798	11	1
the role of	881	13	705	13	0
the expression of	851	14	652	16	-2
the cells were	832	15	665	15	0
we found that	809	16	670	14	2
the effect of	799	17	649	17	0
the number of	791	18	623	19	-1
been shown to	754	19	572	22	-3
of the cell	743	20	603	20	0
for 1 h	732	21	578	21	0
a and b	717	22	636	18	4
the ability of	711	23	568	23	0
for 30 min	709	24	561	24	0
required for the	708	25	502	27	-2
consistent with the	673	26	515	26	0
were incubated with	660	27	526	25	2
as shown in	654	28	463	31	-3
the amount of	607	29	474	29	0
as described in	604	30	498	28	2
activation of the	594	31	448	35	-4
the effects of	593	32	502	27	5
the cell cycle	591	33	421	41	-8
has been shown	579	34	430	40	-6
the cell surface	568	35	468	30	5
is essential for	567	36	459	32	4
involved in the	558	37	456	33	4
expression of the	557	38	448	36	2
at room temperature	555	39	459	32	7
the actin cytoskeleton	551	40	447	37	3
the level of	545	41	446	38	3
the activation of	542	42	407	44	-2
analysis of the	541	43	450	34	9
results indicate that	523	44	342	68	-24
to determine whether	522	45	351	64	-19
were treated with	517	46	420	42	4
we show that	511	47	402	47	0
shown in fig	510	48	350	65	-17
the localization of	510	48	418	43	5
the regulation of	503	49	433	39	10

Appendix 5: Comparison of top 50 3-word patterns: Electrical Engineering

3-word patterns in Electrical Engineering	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
shown in fig	5,574	1	1954	1	0
due to the	2,812	2	1048	2	0
as shown in	2,218	3	771	4	-1
a function of	2,043	4	788	3	1
in fig 1	1,971	5	554	9	-4
in fig 2	1,876	6	547	10	-4
as a function	1,844	7	706	5	2
in order to	1,827	8	646	6	2
the number of	1,699	9	637	7	2
is shown in	1,685	10	577	8	2
in fig 3	1,468	11	486	11	0
based on the	1,400	12	449	12	0
are shown in	1,276	13	449	12	1
as well as	1,262	14	403	16	-2
of the proposed	1,219	15	436	14	1
the case of	1,145	16	430	15	1
with respect to	1,121	17	446	13	4
in fig 4	1,067	18	392	17	1
function of the	1,015	19	354	20	-1
the presence of	985	20	301	28	-8
it can be	978	21	384	18	3
the use of	976	22	323	25	-3
in this letter	971	23	150	90	-67
in terms of	938	24	288	29	-5
the effect of	887	25	326	23	2
in the case	869	26	359	19	7
is given by	859	27	344	22	5
dependence of the	856	28	324	24	4
can be seen	821	29	317	26	3
that of the	814	30	256	37	-7
one of the	805	31	271	33	-2
in this case	785	32	286	30	2
in fig 5	746	33	305	27	6
can be used	729	34	262	34	0
used in the	728	35	288	29	6
at room temperature	727	36	213	48	-12
is used to	723	37	261	35	2
on the other	709	38	276	31	7
in this paper	701	39	346	21	18
may be more	699	40	286	30	10
can be obtained	698	41	260	36	5
value of the	698	41	272	32	9
be more efficient	694	42	200	54	-12
using a second	693	43	202	53	-10
are available with	692	44	200	54	-10
efficient using a	691	45	200	54	-9
more efficient using	691	45	200	54	-9
the value of	665	46	240	40	6
the other hand	662	47	251	38	9
according to the	648	48	234	42	6

Appendix 6: Comparison of top 50 3-word patterns: Mechanical Engineering

3-word patterns in Mechanical Engineering	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
shown in fig	3862	1	2001	1	0
in order to	2273	2	1209	2	0
as shown in	1714	3	834	6	-3
the number of	1684	4	946	3	1
due to the	1648	5	901	4	1
based on the	1609	6	875	5	1
the effect of	1489	7	800	7	0
with respect to	1486	8	738	8	0
the case of	1254	9	617	9	0
in terms of	1174	10	585	10	0
in this paper	1119	11	546	12	-1
as well as	1107	12	553	11	1
a function of	1036	13	544	13	0
it can be	972	14	513	14	0
is shown in	947	15	497	16	-1
are shown in	906	16	446	22	-6
in fig 1	904	17	488	18	-1
in the case	892	18	455	20	-2
the use of	860	19	493	17	2
in fig 2	858	20	427	24	-4
in the present	819	21	484	19	2
as a function	813	22	419	26	-4
the effects of	783	23	499	15	8
values of the	771	24	420	25	-1
respect to the	771	24	402	30	-6
in fig 4	758	25	388	31	-6
one of the	747	26	414	29	-3
is given by	741	27	415	28	-1
in fig 3	736	28	362	36	-8
part of the	733	29	436	23	6
in this study	730	30	452	21	9
according to the	694	31	346	38	-7
1 and 2	693	32	315	45	-13
used in the	684	33	370	33	0
to determine the	677	34	370	33	1
the influence of	672	35	416	27	8
can be obtained	667	36	337	41	-5
in this case	656	37	338	40	-3
the results of	653	38	376	32	6
can be used	649	39	308	49	-10
value of the	639	40	333	42	-2
can be seen	635	41	370	33	8
analysis of the	631	42	365	35	7
the value of	625	43	340	39	4
length of the	621	44	327	43	1
in table 1	611	45	318	44	1
in fig 5	600	46	304	50	-4
function of the	595	47	311	47	0
is used to	568	48	293	53	-5
of the flow	566	49	361	37	12

Appendix 7: Comparison of top 50 3-word patterns: Nuclear Physics

3-word patterns in Nuclear Physics	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
in order to	2,270	1	1223	1	0
a function of	2,234	2	1021	6	-4
due to the	2,231	3	1205	3	0
as a function	2,100	4	941	7	-3
in terms of	1,996	5	1214	2	3
the case of	1,778	6	1026	5	1
the number of	1,673	7	931	8	-1
is given by	1,571	8	1047	4	4
shown in fig	1,515	9	764	12	-3
as well as	1,499	10	834	9	1
part of the	1,374	11	738	13	-2
in the case	1,361	12	784	11	1
with respect to	1,337	13	792	10	3
function of the	1,224	14	584	19	-5
dependence of the	1,127	15	647	14	1
the fact that	1,107	16	647	14	2
values of the	1,074	17	633	16	1
in this case	1,063	18	643	15	3
the ground state	1,060	19	478	28	-9
in fig 1	1,053	20	548	23	-3
on the other	1,032	21	567	21	0
in the present	1,019	22	590	17	5
based on the	996	23	507	26	-3
the other hand	990	24	555	22	2
terms of the	951	25	587	18	7
in fig 2	950	26	400	35	-9
the results of	924	27	472	29	-2
the cross section	910	28	462	30	-2
in this paper	898	29	573	20	9
value of the	878	30	524	24	6
one of the	868	31	432	34	-3
the value of	849	32	498	27	5
energy of the	842	33	390	36	-3
the presence of	840	34	514	25	9
are shown in	813	35	383	39	-4
related to the	789	36	447	32	4
respect to the	778	37	432	34	3
in fig 3	769	38	348	43	-5
the effect of	762	39	437	33	6
agreement with the	752	40	381	40	0
given by the	732	41	447	31	10
the experimental data	727	42	390	36	6
because of the	688	43	304	63	-20
is shown in	686	44	343	46	-2
note that the	678	45	389	37	8
of the two	675	46	329	52	-6
corresponding to the	669	47	390	36	11
degrees of freedom	664	48	338	48	0
analysis of the	640	49	333	50	-1
there is no	634	50	386	38	12

Appendix 8: Comparison of top 50 3-word patterns: Economics

3-word patterns in Economics	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
the number of	1,910	1	2147	1	0
the united states	1,416	2	1484	2	0
there is a	1,334	3	1355	4	-1
the effect of	1,251	4	1369	3	1
the fact that	1,218	5	1315	5	0
with respect to	1,147	6	1203	7	-1
the set of	1,107	7	1110	8	-1
the value of	1,105	8	1205	6	2
there is no	1,028	9	1063	10	-1
in which the	983	10	1046	11	-1
in this case	967	11	1020	12	-1
as well as	947	12	1077	9	3
in terms of	934	13	986	15	-2
in the united	930	14	988	13	1
the effects of	910	15	987	14	1
an increase in	906	16	974	16	0
in this paper	884	17	966	17	0
in order to	859	18	925	19	-1
is given by	857	19	863	22	-3
we assume that	843	20	926	18	2
the level of	810	21	854	23	-2
the case of	803	22	888	20	2
the distribution of	780	23	868	21	2
increase in the	775	24	831	24	0
of the model	771	25	818	25	0
on the other	735	26	756	29	-3
value of the	709	27	754	30	-3
the size of	704	28	784	26	2
if and only	700	29	702	35	-6
and only if	698	30	700	36	-6
is that the	695	31	760	28	3
the impact of	676	32	743	31	1
the probability of	657	33	699	37	-4
changes in the	655	34	709	34	0
a number of	652	35	724	32	3
the presence of	650	36	710	33	3
the other hand	649	37	667	40	-3
that there is	630	38	656	42	-4
as a result	609	39	662	41	-2
the rate of	592	40	622	46	-6
the existence of	589	41	593	51	-10
in the first	588	42	672	39	3
the cost of	580	43	615	47	-4
percent of the	578	44	764	27	17
in the case	566	45	609	49	-4
size of the	564	46	640	43	3
assume that the	562	47	603	50	-3
consistent with the	558	48	610	48	0
likely to be	558	48	623	45	3
the u s	552	49	609	49	0

Appendix 9: Comparison of top 50 3-word patterns: Marketing

3-word patterns in Marketing	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
the number of	1585	1	1586	1	0
as well as	1573	2	1538	2	0
the effect of	1441	3	1481	3	0
the effects of	1413	4	1431	4	0
in terms of	1399	5	1339	7	-2
the impact of	1355	6	1356	5	1
the use of	1313	7	1352	6	1
the role of	1167	8	1122	9	-1
the basis of	1093	9	1127	8	1
on the basis	1033	10	1068	10	0
one of the	935	11	918	11	0
the relationship between	919	12	908	12	0
the importance of	905	13	883	13	0
each of the	852	14	874	14	0
the level of	832	15	760	17	-2
more likely to	808	16	800	15	1
likely to be	779	17	779	16	1
based on the	733	18	705	23	-5
there is a	732	19	732	20	-1
are likely to	729	20	752	19	1
extent to which	724	21	755	18	3
the nature of	710	22	709	21	1
the extent to	702	23	732	20	3
in which the	680	24	708	22	2
is likely to	678	25	680	24	1
the context of	671	26	666	26	0
the results of	669	27	637	27	0
the development of	668	28	580	34	-6
with respect to	644	29	673	25	4
as a result	639	30	624	30	0
may not be	637	31	637	27	4
the case of	636	32	634	28	4
a number of	624	33	572	36	-3
in this study	612	34	596	31	3
consistent with the	594	35	628	29	6
in other words	578	36	587	32	4
in the context	574	37	573	35	2
in order to	570	38	466	50	-12
the influence of	570	58	572	36	22
because of the	551	39	583	33	6
a function of	550	40	563	38	2
of market orientation	548	41	458	51	-10
of the product	538	42	567	37	5
nature of the	505	43	510	40	3
related to the	504	44	491	44	0
some of the	503	45	515	39	6
in addition to	501	46	502	42	4
part of the	488	47	489	46	1
the value of	484	48	499	43	5
there is no	477	49	490	45	4

Appendix 10: Comparison of top 50 3-word patterns: Philosophy

3-word patterns in Philosophy	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
that it is	2,660	1	2468	1	0
it is not	2,304	2	2096	2	0
the fact that	2,245	3	2060	3	0
there is a	2,211	4	2006	6	-2
there is no	2,205	5	2051	4	1
that there is	2,121	6	2032	5	1
in terms of	1,933	7	1791	7	0
in order to	1,613	8	1499	8	0
to say that	1,366	9	1254	9	0
it is a	1,184	10	1088	12	-2
that there are	1,175	11	1107	11	0
with respect to	1,155	12	1165	10	2
the existence of	1,154	13	1004	17	-4
the concept of	1,148	14	1064	13	1
the case of	1,100	15	1033	14	1
to be a	1,068	16	1010	16	0
is not a	1,053	17	955	19	-2
the claim that	1,052	18	1016	15	3
the notion of	1,037	19	949	20	-1
i do not	1,012	20	963	18	2
on the other	1,011	21	917	21	0
it does not	971	22	893	22	0
the possibility of	949	23	856	25	-2
the nature of	940	24	886	23	1
in the case	917	25	860	24	1
in which the	901	26	840	26	0
is to be	890	27	827	27	0
one of the	867	28	787	30	-2
as well as	853	29	769	34	-5
part of the	837	30	820	28	2
it would be	835	31	811	29	2
is that the	829	32	781	31	1
the basis of	827	33	780	32	1
it is the	808	34	746	36	-2
in the sense	807	35	771	33	2
a matter of	806	36	752	35	1
but it is	796	37	703	42	-5
is not the	795	38	718	40	-2
the problem of	789	39	667	45	-6
if it is	788	40	735	38	2
the idea that	785	41	721	39	2
of the world	759	42	703	42	0
to show that	753	43	707	41	2
be able to	742	44	740	37	7
the idea of	738	45	670	44	1
the other hand	734	46	670	44	2
on the basis	722	47	679	43	4
to be the	711	48	655	47	1
and so on	692	49	628	50	-1
this is not	686	50	619	51	-1

Appendix 11: Comparison of top 50 3-word patterns: Sociology

3-word patterns in Sociology	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
in terms of	1,757	1	1784	1	0
as well as	1,717	2	1766	2	0
the united states	1,227	3	1475	3	0
in order to	1,149	4	1123	7	-3
one of the	1,144	5	1124	6	-1
a number of	989	6	1011	8	-2
the role of	945	7	952	10	-3
the number of	918	8	1129	4	4
the effects of	910	9	1125	5	4
the fact that	901	10	878	14	-4
part of the	886	11	870	15	-4
there is a	863	12	891	13	-1
the development of	857	13	920	11	2
in the united	801	14	965	9	5
on the other	767	15	768	19	-4
more likely to	735	16	864	16	0
the importance of	733	17	782	17	0
of the social	727	18	686	23	-5
the case of	724	19	769	18	1
the use of	720	20	751	21	-1
the concept of	709	21	657	29	-8
some of the	689	22	670	26	-4
the effect of	685	23	918	12	11
it is not	683	24	671	25	-1
at the same	676	25	683	24	1
the relationship between	675	26	741	22	4
there is no	675	26	683	24	2
the same time	674	27	668	27	0
in relation to	645	28	589	31	-3
the impact of	644	29	755	20	9
in which the	624	30	655	30	0
that it is	595	31	575	36	-5
likely to be	589	32	663	28	4
the nature of	589	32	581	34	-2
ways in which	584	33	544	41	-8
such as the	573	34	586	33	1
the basis of	572	35	554	39	-4
the context of	558	36	530	42	-6
as a result	551	37	580	35	2
the end of	549	38	547	40	-2
the level of	541	39	588	32	7
the other hand	531	40	526	43	-3
the study of	528	41	564	38	3
the idea of	521	42	500	47	-5
in other words	514	43	525	44	-1
in the case	502	44	518	45	-1
the emergence of	484	45	568	37	8
to be a	468	46	459	52	-6
a variety of	463	47	499	48	-1
of the world	462	48	444	55	-7

Appendix 12: Comparison of top 50 4-word patterns: Cell Biology

4-word patterns in Cell Biology	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank diff.
in the presence of	PP+OF	1970	1	1627	1	0
in the absence of	PP+OF	1359	2	1091	2	0
has been shown to	V/A+TO	400	3	299	5	-2
cells were transfected with	PV+PP	396	4	334	3	1
to the plasma membrane	OPP	364	5	310	4	1
as shown in fig	ADV	353	6	233	8	-2
these results indicate that	VP	349	7	226	12	-5
is required for the	BE+NP/AP	342	8	242	7	1
the presence or absence	NP	295	9	231	9	0
presence or absence of	NP+OF	294	10	229	10	0
it is possible that	IT+AP/VP	292	11	227	11	0
these results suggest that	VP	285	12	253	6	6
in the presence or	PP+OF	281	13	222	14	-1
the presence of the	NP+OF	274	14	225	13	1
in materials and methods	OPP	264	15	214	15	0
in the regulation of	PP+OF	259	16	225	13	3
described in materials and	PV+PP	248	17	199	17	0
for 1 h at	OPP	248	17	210	16	1
as described in materials	ADV	246	18	197	18	0
cells were treated with	PV+PP	246	18	197	18	0
as shown in figure	ADV	245	19	184	21	-2
according to the manufacturer's	OPP	234	20	184	21	-1
for 30 min at	OPP	232	21	193	19	2
see materials and methods	VP	228	22	197	18	4
had no effect on	VP	227	23	184	21	2
as well as the	OTHER	226	24	185	20	4
at the plasma membrane	OPP	224	25	166	28	-3
at 37 c for	OPP	223	26	185	20	6
have been shown to	V/A+TO	222	27	171	25	2
is consistent with the	BE+NP/AP	219	28	151	33	-5
similar results were obtained	VP	217	29	179	22	7
in the case of	PP+OF	212	30	177	23	7
on the basis of	PP+OF	211	31	129	46	-15
was performed as described	PV+PP	211	31	171	25	6
been shown to be	PV+PP	209	32	150	34	-2
mm hepes ph 7	NP	209	32	156	31	1
of the actin cytoskeleton	OPP	208	33	168	27	6
cells were incubated with	PV+PP	204	34	154	32	2
here we show that	VP	202	35	145	37	-2
min at 37 c	NP+PP	199	36	170	26	10
h at 4 c	OPP	195	37	149	35	2
min at 4 c	NP+PP	193	38	164	29	9
these data suggest that	VP	191	39	172	24	15
play a role in	VP	185	40	159	30	10
of the cell cycle	OPP	183	41	136	40	1
we have shown that	VP	180	42	126	47	-5
is not required for	BE+NP/AP	179	43	129	46	-3
was added to the	PV+PP	178	44	139	39	5
in this study we	OPP	176	45	146	36	9
on the other hand	OPP	176	45	144	38	7

Appendix 13: Comparison of top 50 4-word patterns: Electrical Engineering

4-word patterns in Electrical Engineering	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank diff.
as a function of	ADV	1837	1	704	1	0
as shown in fig	ADV	1750	2	608	2	0
is shown in fig	BE+NP/AP	1484	3	522	3	0
shown in fig 1	PV+PP	1039	4	294	6	-2
shown in fig 2	PV+PP	885	5	245	10	-5
are shown in fig	PV+PP	870	6	300	5	1
in the case of	PP+OF	810	7	333	4	3
a function of the	NP+OF	765	8	283	7	1
shown in fig 3	PV+PP	691	9	224	11	-2
on the other hand	OPP	660	10	251	8	2
with respect to the	OPP	638	11	249	9	2
in this letter we	OPP	590	12	-	-	-
shown in fig 4	PV+PP	512	13	188	12	1
in fig 2 the	OPP	390	14	121	16	-2
as well as the	OTHER	383	15	113	19	-4
it can be seen	IT+AP/VP	381	16	148	14	2
shown in fig 5	PV+PP	363	17	155	13	4
the performance of the	NP+OF	360	18	101	25	-7
the thickness of the	NP+OF	355	19	101	25	-6
can be used to	PV+PP	354	20	133	15	5
can be expressed as	PV+PP	352	21	117	17	4
in fig 1 the	OPP	350	22	-	-	-
in this case the	OPP	316	23	112	20	3
is due to the	BE+NP/AP	316	24	121	16	8
can be seen that	VP+THAT	296	25	116	18	7
in terms of the	PP+OF	288	26	-	-	-
the size of the	NP+OF	287	27	108	24	3
it is possible to	IT+AP/VP	279	28	111	21	7
than that of the	OTHER	279	28	-	-	-
in fig 3 the	OPP	278	29	-	-	-
fig 2 shows the	VP	267	30	-	-	-
shown in fig 7	PV+PP	266	31	110	22	9
as can be seen	ADV	265	32	108	24	8
is based on the	PV+PP	261	33	-	-	-
the fact that the	NP+THAT	261	33	-	-	-
the length of the	NP+OF	260	34	-	-	-
the effect of the	NP+OF	259	35	111	21	14
in the range of	PP+OF	258	36	-	-	-
the case of the	NP+OF	253	37	-	-	-
it should be noted	IT+AP/VP	250	38	-	-	-
the output of the	NP+OF	245	39	-	-	-
shown in fig 6	PV+PP	241	40	101	25	15
the magnitude of the	NP+OF	240	41	-	-	-
is the number of	BE+NP/AP	239	42	-	-	-
can be written as	PV+PP	237	43	-	-	-
the amplitude of the	NP+OF	237	43	109	23	20
are shown in figs	PV+PP	236	44	-	-	-
at the same time	OPP	236	44	-	-	-
fig 3 shows the	VP	230	45	-	-	-
the temperature dependence of	NP+OF	230	45	-	-	-

Appendix 14: Comparison of top 50 4-word patterns: Mech. Eng

4-word patterns in Mechanical Engineering	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank diff.
as shown in fig	ADV	1410	1	696	1	0
is shown in fig	BE+NP/AP	838	2	439	2	0
as a function of	ADV	810	3	419	3	0
in the case of	PP+OF	799	4	404	4	0
with respect to the	OPP	763	5	396	5	0
are shown in fig	PV+PP	616	6	306	6	0
shown in fig 1	PV+PP	552	7	297	7	0
on the other hand	OPP	498	8	270	8	0
shown in fig 2	PV+PP	460	9	228	9	0
in terms of the	PP+OF	392	10	211	11	-1
a function of the	NP+OF	385	11	211	10	1
shown in fig 3	PV+PP	379	12	187	15	-3
as well as the	OTHER	368	13	193	14	-1
the length of the	NP+OF	367	14	195	13	1
shown in fig 4	PV+PP	362	15	183	16	-1
it can be seen	IT+AP/VP	357	16	214	10	6
can be expressed as	PV+PP	353	17	161	23	-6
the effect of the	NP+OF	337	18	207	12	6
can be used to	PV+PP	333	19	160	24	-5
can be written as	PV+PP	297	20	160	24	-4
shown in fig 5	PV+PP	297	20	144	28	-8
on the basis of	PP+OF	291	21	155	26	-5
the magnitude of the	NP+OF	291	21	167	21	0
it is possible to	IT+AP/VP	290	22	172	18	4
for the case of	PP+OF	287	23	129	35	-12
can be seen that	VP+THAT	283	24	165	22	2
the position of the	NP+OF	278	25	137	31	-6
it should be noted	IT+AP/VP	276	26	170	19	7
the size of the	NP+OF	274	27	172	18	9
in the present study	OPP	272	28	168	20	8
the influence of the	NP+OF	268	29	174	17	12
the end of the	NP+OF	267	30	161	23	7
at the end of	PP+OF	265	31	146	27	4
is based on the	PV+PP	265	31	132	33	-2
it is necessary to	IT+AP/VP	258	32	128	36	-4
should be noted that	VP+THAT	258	33	159	25	8
shown in fig 6	PV+PP	256	34	139	30	4
is assumed to be	PV+PP	251	35	131	34	1
the center of the	NP+OF	242	36	137	31	5
the results of the	NP+OF	239	37	131	34	3
in this case the	OPP	238	38	128	36	2
the shape of the	NP+OF	232	39	133	32	7
in this paper the	OPP	231	40	106	45	-5
the direction of the	NP+OF	227	41	142	29	12
the accuracy of the	NP+OF	224	42	142	29	13
in the form of	PP+OF	223	43	114	41	2
the value of the	NP+OF	212	44	117	39	5
the surface of the	NP+OF	211	45	116	40	5
the fact that the	NP+THAT	209	46	109	43	3
is the number of	BE+NP/AP	206	47	114	41	6

Appendix 15: Comparison of top 50 4-word patterns: Nuclear Physics

4-word patterns in Nuclear Physics	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank diff.
as a function of	ADV	2092	1	939	1	0
in the case of	PP+OF	1182	2	662	2	0
on the other hand	OPP	983	3	554	3	0
a function of the	NP+OF	869	4	379	6	-2
in terms of the	PP+OF	860	5	535	4	1
with respect to the	OPP	757	6	416	5	1
the fact that the	NP+THAT	508	7	300	7	0
as well as the	OTHER	475	8	266	8	0
in this paper we	OPP	469	9	300	7	2
the case of the	NP+OF	460	10	262	9	1
is shown in fig	BE+NP/AP	427	11	213	12	-1
are shown in fig	PV+PP	403	12	202	15	-3
can be written as	PV+PP	397	13	248	11	2
in this case the	OPP	397	14	250	10	4
shown in fig 1	PV+PP	345	15	166	20	-5
is given by the	BE+NP/AP	333	16	210	13	3
by a factor of	PP+OF	317	17	148	29	-12
of the order of	PP+OF	317	17	162	22	-5
the other hand the	NP	315	18	177	18	0
the value of the	NP+OF	313	19	209	14	5
the results of the	NP+OF	309	20	153	26	-6
the effect of the	NP+OF	304	21	158	24	-3
the energy of the	NP+OF	303	22	139	33	-11
in the framework of	PP+OF	302	23	173	19	4
the ratio of the	NP+OF	292	24	143	31	-7
can be found in	PV+PP	288	25	165	21	4
in this section we	OPP	287	26	195	16	10
shown in fig 2	PV+PP	286	27	120	46	-19
the contribution of the	NP+OF	284	28	146	30	-2
the ground state of	NP+OF	277	29	114	49	-20
is due to the	BE+NP/AP	275	30	146	30	0
the sum of the	NP+OF	274	31	162	22	9
in the present work	OPP	270	32	160	23	9
is consistent with the	BE+NP/AP	267	33	151	28	5
the size of the	NP+OF	266	34	152	27	7
in the presence of	PP+OF	257	35	184	17	18
the mass of the	NP+OF	249	36	140	32	4
as shown in fig	ADV	247	37	133	37	0
the calculation of the	NP+OF	246	38	155	25	13
the values of the	NP+OF	245	39	133	37	2
it is possible to	IT+AP/VP	241	40	129	39	1
to the fact that	PP+THAT	240	41	127	42	-1
shown in fig 3	PV+PP	238	42	108	54	-12
as can be seen	ADV	235	43	123	45	-2
in addition to the	OPP	234	44	120	46	-2
is related to the	BE+NP/AP	233	45	133	37	8
we find that the	VP+THAT	233	45	128	41	4
for the case of	PP+OF	232	46	129	39	7
for the first time	OPP	229	47	-	-	-
the analysis of the	NP+OF	229	47	103	56	-9

Appendix 16: Comparison of top 50 4-word patterns: Economics

4-word patterns in Economics	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
in the united states	OPP	735	1	787	1	0
if and only if	OTHER	682	2	697	2	0
on the other hand	OPP	641	3	666	3	0
the size of the	NP+OF	474	4	532	4	0
in the absence of	PP+OF	409	5	436	6	-1
in the case of	PP+OF	408	6	445	5	1
an increase in the	NP+PP	380	7	405	7	0
as a function of	ADV	371	8	384	8	0
is an element of	OTHER	324	9	159	54	-45
with respect to the	OPP	309	10	312	12	-2
the fact that the	NP+THAT	292	11	319	11	0
in this case the	OPP	286	12	297	14	-2
on the basis of	PP+OF	282	13	325	10	3
it is easy to	IT+AP/VP	280	14	284	18	-4
the value of the	NP+OF	280	14	300	13	1
in terms of the	PP+OF	278	15	286	16	-1
in this paper we	OPP	273	16	300	13	3
in this section we	OPP	268	17	291	15	2
in the presence of	PP+OF	266	18	273	21	-3
in the sense that	PP+THAT	255	19	255	24	-5
we assume that the	VP+THAT	254	20	276	18	2
in the context of	PP+OF	246	21	274	20	1
at the beginning of	PP+OF	245	22	284	18	4
at the end of	PP+OF	245	22	285	17	5
as well as the	OTHER	235	23	276	19	4
the case in which	NP	231	24	257	23	1
the rest of the	NP+OF	229	25	267	22	3
at the same time	OPP	226	26	250	25	1
is consistent with the	BE+NP/AP	219	27	231	27	0
the united states and	NP	214	28	214	32	-4
the end of the	NP+OF	210	29	239	26	3
that there is a	THAT	207	30	209	33	-3
the effect of the	NP+OF	207	30	228	28	2
as a result of	ADV	203	31	218	31	0
the change in the	NP	196	32	221	30	2
in the long run	OPP	194	33	227	29	4
federal reserve bank of	NP+OF	191	34	201	35	-1
in the united kingdom	OPP	190	35	196	37	-2
the extent to which	NP	188	36	204	34	2
the existence of a	NP+OF	180	37	170	50	-13
a function of the	NP+OF	178	38	184	42	-4
to the extent that	PP+THAT	177	39	209	33	6
the growth rate of	NP+OF	176	40	178	46	-6
can be written as	PV+PP	171	41	187	40	1
is likely to be	BE+NP/AP	171	41	193	38	3
is equal to the	BE+NP/AP	169	42	161	52	-10
it is possible to	IT+AP/VP	167	43	179	45	-2
are more likely to	V/A+TO	165	44	187	39	5
the sum of the	NP+OF	163	45	172	48	-3
a large number of	NP+OF	162	46	172	48	-2

Appendix 17: Comparison of top 50 4-word patterns: Marketing

4-word patterns in Marketing	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
on the basis of	PP+OF	1024	1	1062	1	0
the extent to which	NP	701	2	731	2	0
in the context of	PP+OF	561	3	562	3	0
the degree to which	NP	400	4	406	4	0
in the case of	PP+OF	390	5	380	5	0
as well as the	OTHER	362	6	349	7	-1
on the other hand	OPP	353	7	331	9	-2
are more likely to	V/A+TO	348	8	350	6	2
it is important to	IT+AP/VP	329	9	334	8	1
as a function of	ADV	303	10	314	10	0
the nature of the	NP+OF	303	10	303	12	-2
the basis of the	NP+OF	293	11	301	13	-2
for each of the	PP+OF	280	12	287	14	-2
as a result of	ADV	274	13	265	16	-3
in the united states	OPP	263	14	309	11	3
in terms of the	PP+OF	254	15	242	18	-3
are likely to be	V/A+TO	247	16	250	17	-1
at the same time	OPP	241	17	224	20	-3
than or equal to	OTHER	239	18	274	15	3
the results of the	NP+OF	235	19	220	21	-2
is likely to be	BE+NP/AP	230	20	229	19	1
in the absence of	PP+OF	206	21	204	22	-1
can be used to	PV+PP	204	22	203	23	-1
to the extent that	PP+THAT	195	23	198	24	-1
in the form of	PP+OF	194	24	196	25	-1
the size of the	NP+OF	193	25	187	27	-2
with respect to the	OPP	192	26	196	25	1
the quality of the	NP+OF	189	27	220	21	6
extent to which the	NP	188	28	190	26	2
is consistent with the	BE+NP/AP	178	29	187	27	2
a function of the	NP+OF	174	30	174	28	2
at the time of	PP+OF	170	31	146	35	-4
is more likely to	BE+NP/AP	168	32	166	32	0
more likely to be	OTHER	167	33	168	30	3
in addition to the	OPP	163	34	167	31	3
it is possible that	IT+AP/VP	159	35	172	29	6
on the part of	PP+OF	149	36	152	34	2
a high degree of	NP+OF	147	37	133	42	-5
is based on the	PV+PP	147	37	143	37	0
less than or equal	OTHER	147	37	162	33	4
the total number of	NP+OF	145	38	152	34	4
a large number of	NP+OF	143	39	140	38	1
in the next section	OPP	141	40	152	34	6
in the presence of	PP+OF	141	40	138	39	1
the relationship between the	NP	140	41	146	36	5
a wide range of	NP+OF	136	42	137	40	2
a high level of	NP+OF	133	43	130	44	-1
of the variance in	OPP	131	44	126	45	-1
the impact of the	NP+OF	131	44	131	43	1
the case of the	NP+OF	128	45	120	49	-4

Appendix 18:Comparison of top 50 4-word patterns: Philosophy

4-word patterns in Philosophy	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
in the case of	PP+OF	826	1	774	1	0
on the other hand	OPP	730	2	667	2	0
on the basis of	PP+OF	701	3	664	3	0
that there is a	THAT	678	4	643	4	0
if and only if	OTHER	545	5	495	5	0
that there is no	THAT	499	6	486	6	0
in terms of the	PP+OF	453	7	424	7	0
at the same time	OPP	427	8	412	8	0
in the sense that	PP+THAT	385	9	360	9	0
the fact that the	NP+THAT	362	10	338	10	0
on the one hand	OPP	352	11	316	11	0
it is true that	IT+AP/VP	339	12	296	12	0
in such a way	OPP	311	13	270	16	-3
that it is not	THAT	302	14	273	15	-1
in so far as	OPP	295	15	239	23	-8
the way in which	NP	292	16	282	14	2
in a way that	PP+THAT	286	17	286	13	4
in the context of	PP+OF	280	18	264	17	1
in the sense of	PP+OF	270	19	257	18	1
as we have seen	ADV	259	20	251	21	-1
it seems to me	IT+AP/VP	256	21	254	19	2
to be able to	VP	251	22	253	20	2
for the sake of	OPP	247	23	249	22	1
as a matter of	ADV	244	24	212	27	-3
from the fact that	PP+THAT	233	25	208	29	-4
that it is a	THAT	233	25	213	26	-1
it is possible to	IT+AP/VP	232	26	212	27	-1
to say that the	VP	232	26	224	25	1
that is to say	THAT	229	27	207	30	-3
with respect to the	OPP	227	28	237	24	4
there is no reason	PRON	226	29	198	32	-3
the property of being	NP+OF	223	30	188	37	-7
such a way that	OTHER	222	31	190	35	-4
the nature of the	NP+OF	215	32	210	28	4
it is not the	OTHER	212	33	170	45	-12
of the concept of	PP+OF	212	33	186	39	-6
would have to be	OTHER	205	34	197	33	1
the notion of a	NP+OF	204	35	187	38	-3
in the philosophy of	PP+OF	203	36	206	31	5
it is important to	IT+AP/VP	202	37	195	34	3
but it is not	IT+AP/VP	200	38	162	50	-12
in a position to	OPP	200	38	239	23	15
in the same way	OPP	199	39	188	37	2
it is clear that	IT+AP/VP	199	40	184	40	0
it is hard to	IT+AP/VP	198	41	179	42	-1
in the light of	PP+OF	197	42	189	36	6
one and the same	NP	194	43	177	43	0
in the absence of	PP+OF	189	44	167	48	-4
the case of the	NP+OF	188	45	182	41	4
in the course of	PP+OF	186	46	170	45	1

Appendix 19: Comparison of top 50 4-word patterns: Sociology

4-word patterns in Sociology	Form	Isolexical frequency	Isolexical rank	Isotextual frequency	Isotextual rank	Rank difference
in the united states	OPP	767	1	767	1	0
at the same time	OPP	640	2	632	2	0
on the other hand	OPP	531	3	525	3	0
in the case of	PP+OF	474	4	486	4	0
in the context of	PP+OF	439	5	421	6	-1
on the basis of	PP+OF	436	6	428	5	1
the ways in which	NP	391	7	355	10	-3
the extent to which	NP	385	8	388	7	1
in terms of the	PP+OF	357	9	371	8	1
on the one hand	OPP	342	10	338	11	-1
are more likely to	V/A+TO	307	11	361	9	2
the end of the	NP+OF	304	12	296	15	-3
as well as the	OTHER	302	13	312	12	1
as a result of	ADV	297	14	301	13	1
it is important to	IT+AP/VP	274	15	299	14	1
in the form of	PP+OF	249	16	250	17	-1
to the extent that	PP+THAT	243	17	265	16	1
one of the most	NP	236	18	235	18	0
the nature of the	NP+OF	231	19	234	19	0
the way in which	NP	220	20	198	22	-2
at the level of	PP+OF	218	21	203	21	0
more likely to be	OTHER	204	22	224	20	2
in relation to the	OPP	197	23	175	26	-3
at the end of	PP+OF	179	24	189	23	1
the social construction of	NP+OF	176	25	170	27	-2
the role of the	NP+OF	175	26	166	29	-3
the case of the	NP+OF	173	27	181	24	3
on the part of	PP+OF	172	28	154	35	-7
the fact that the	NP+THAT	169	29	169	28	1
of the social sciences	OPP	168	30	146	46	-16
in the face of	PP+OF	164	31	159	32	-1
of women in the	OPP	157	32	156	34	-2
at the time of	PP+OF	150	33	160	31	2
in the absence of	PP+OF	149	34	161	30	4
it is possible to	IT+AP/VP	149	34	140	38	-4
in the sense that	PP+THAT	148	35	144	37	-2
there has been a	PRON	145	36	137	40	-4
is one of the	BE+NP/AP	143	37	131	44	-7
the degree to which	NP	142	38	158	33	5
the united states and	NP	139	39	177	25	14
that there is a	THAT	134	40	140	38	2
of the welfare state	OPP	133	41	123	49	-8
in the social sciences	OPP	132	42	134	42	0
as a means of	ADV	131	43	146	36	7
can be seen as	VP+ADV	131	43	120	52	-9
in the course of	PP+OF	131	43	127	47	-4
in the process of	PP+OF	130	44	121	51	-7
as part of the	ADV	129	45	122	50	-5
in the number of	PP+OF	127	46	134	42	4
a wide range of	NP+OF	126	47	137	40	7

Appendix 20: Lexis framed by *in the case of the*

20a) Science disciplines

Cell Bio	aggresomes; cells; channels; column; cones; CSP83-198; E2; hammerhead; ligament(s); markers; melanosome; mutant; panel; promoter; protein; receptor; regulator; replisome; Spindles; state; strain; strengths; transgene; treatments; TTSS; variant
Electrical Engineering	absence; anisotropy; components; control; converter (2); devices; dimension; electrode; etching; fault; field; film (4); films (3); gas (2) generation; grid (3); H-bridge; heterostructure; HF-UPC; HPA; Hugoniot; igniter; invertor; ions; loop; machine (2); material; parameter; paths; range (2) rectifier; rectifiers; resonator; sample (3); samples (4); set; signal; spacing (2); strategy; structure (2); system (3); topology; transmission (2); voltages (3)
Mechanical Engineering	analysis; beams; behaviour; bending; bodies; boundary; capacity (4); cascade; curve (3); CVT; data (3); diameter; drive (5); example; features; film; fluorosopy; fraction; fracture; functions; gap (2); generation; IKM; industry; interface; lubricant; manipulator; material; mechanism (3); mechanisms (2); method; model (5); modulus; nozzle; number; onset; particle (2); particles (2); perturbation (2); platform; pressure; problem (2); prototype; punch; results; set (2); shortness; singularity; size; steel; surface; system; tool (3); Venturi; wake (2); waves (2)
Nuclear Physics	0nbb; 1); analysis; anti-neutrinos; antiproton; background (4); bang; brane; calculation; capture (2); chain; channel (3); classification; compactification; constraints; contribution; coupling; couplings; data; decay; decays; description; deuteron (3); diagrams; dibaryon; doublets; DR (2); droplet; effects; elements; emission (2); energy (2); EOS; equation; excitation (2); experiment; factor (2); formulation; function (3); functions; g; galaxies; gas; GMR; Hamiltonian; IBFBPM; instanton; interaction (3); isomer (2) isotones; isotopes (3); Lagrangian; lattice; level; levels; map; measurement (2); measurements; method; mode; model (13); models; MSSM; necessity; nuclei (4); nucleon (2) nucleus (6); operator; operators; oscillator; particle (2); phase (3); plane; potential (4); process (2); propagator; quarks (2); rate; reaction (10); reactions (4); realisation; reduction; region; resonance (2); result; scattering; section (4); self-energy; setup (2); SGII (3); SIII (5); simulation; situation; solution; states (5); supposition; system (4); target; theory (2); torus; transition (3); tube; wall; waves

20b) Non-Science disciplines

Economics	agents; caste; components; country; crisis; criterion (2); cut; dictators; dummy; economies; economy (5); equation; example; formula (2); function (2); games; IMF; industry (2); Kingdom (3); L-1011; logit; mark; miracle; model (6); Netherlands (2); paper; parameters; period; probability; problem (2); programme; rate; regressors (3); results; score; security; States (2); structure; survey; tournament; treatments (2); Treaty (2); vaccine; yen
Marketing	advertisement (2); arrangement; attention; brand (11); categories (3); category (3); choice; competition; computer (2); conditions; conference; congress; designs; elasticity; emphasis (2); entrant; example (3); extension (3); fee; former; index; industry (2); latter (2); market (3); merger; model (5); models; name (2); parties; permit; provider; questions; relationships; scale; scandal; specification; sponsorship; state (2); studies; study; Taurus; variable; version
Philosophy	agent; agents; A-image (5); assassin; attitudes; attribution; B-image (4); blasts; body; brightness-line; brightness-space; builders; categories (4); causation; child; community (2); concept (2); concepts; connectives; constants; copula; difficulties; disclosure; distinction (2); doctrine; dog (2); driver; embryo; endorsement; engine (2); entities; equations; essence (4); event; example (2); exemplification; existence; experiment; eye; fish; form; former; formulation; game; hologram-viewer; hypothesis; I; 'It's raining'; laws; liar; lines, waves; lotteries; moment; moral, legal; names; number; numbers; nurse; occurrent; painter; pendulum; person (2); perspective; philosophy; pigeon; Prince; prince, cobbler; principle (2); problem (2); properties (4); proposition (3); psychologist; quantifier; quantifier-prefixes; question (4); R; reals (2); reasoning; reduction; relation (3); relations; representations; revolution; right; sceptic (2); sciences (4); senses; sexers; shoppers; solution; substitutions; suggestion; system; theories; theory (3); Thesis; Tractatus; village; Way; wedding-speech; whole; woodsellers; workplace; yogi-concepts
Sociology	anti-materialism; anti-Americanism; Armenians; automaker; bank (2); BHPS; bikers; bricklayer; committee; commons; communication; company; concepts; contract; control; countries; cult; cultures; data; Deal development; East; elite; environmentalists (2); fa'afafine; feminists; force; forces; former (3); frame; GDR; genocide; girls; greens (4); grid; group; groups; hippies; homeless; industries; industry (3); Internet (3); Israeli-Arabs; Japanese; jewellery; latter (2); LFS; MA; Man; Mardu (2); marketing; measure; media; military; miner and soldier; model; monopoly; movement (5); Muslims; nationalism; navy; NIRA; objects; OCG; officers and servants; owners; panic; patient; Plot; programme; PUH; respondent; Right; schema; School; second; sibling (2); SMOs; States (3); studies; super-panopticon; survey; system; tagging (4); taking up; telephoning; thesis; Trobrianders; UK; USA (3); videoing; weavers; world; Zapatistas