

ACADEMIC VOCABULARY IN LEARNER WRITING

From Extraction
to Analysis

MAGALI PAQUOT

RESEARCH IN CORPUS AND DISCOURSE
SERIES EDITORS: WOLFGANG TEUBERT AND MICHAELA MAHLBERG

B L O O M S B U R Y

Academic Vocabulary in Learner Writing

Corpus and Discourse

Series editors: Wolfgang Teubert, University of Birmingham, and Michaela Mahlberg, University of Liverpool.

Editorial Board: Paul Baker (Lancaster), Frantisek Čermák (Prague), Susan Conrad (Portland), Geoffrey Leech (Lancaster), Dominique Maingueneau (Paris XII), Christian Mair (Freiburg), Alan Partington (Bologna), Elena Tognini-Bonelli (Siena and TWC), Ruth Wodak (Lancaster), Feng Zhiwei (Beijing).

Corpus linguistics provides the methodology to extract meaning from texts. Taking as its starting point the fact that language is not a mirror of reality but lets us share what we know, believe and think about reality, it focuses on language as a social phenomenon, and makes visible the attitudes and beliefs expressed by the members of a discourse community.

Consisting of both spoken and written language, discourse always has historical, social, functional, and regional dimensions. Discourse can be monolingual or multilingual, interconnected by translations. Discourse is where language and social studies meet.

The *Corpus and Discourse* series consists of two strands. The first, *Research in Corpus and Discourse*, features innovative contributions to various aspects of corpus linguistics and a wide range of applications, from language technology via the teaching of a second language to a history of mentalities. The second strand, *Studies in Corpus and Discourse*, is comprised of key texts bridging the gap between social studies and linguistics. Although equally academically rigorous, this strand will be aimed at a wider audience of academics and postgraduate students working in both disciplines.

Research in Corpus and Discourse

Conversation in Context

A Corpus-driven Approach

With a preface by Michael McCarthy

Christoph Rühlemann

Corpus-Based Approaches to English Language Teaching

Edited by Mari Carmen Campoy, Begona Bellés-Fortuno and M^a Lluïsa Gea-Valor

Corpus Linguistics and World Englishes

An Analysis of Xhosa English

Vivian de Klerk

Evaluation and Stance in War News

A Linguistic Analysis of American, British and Italian television news reporting of the 2003 Iraqi war

Edited by Louann Haarman and Linda Lombardo

Evaluation in Media Discourse
Analysis of a Newspaper Corpus
Monika Bednarek

Historical Corpus Stylistics
Media, Technology and Change
Patrick Studer

Idioms and Collocations
Corpus-based Linguistic and Lexicographic Studies
Edited by Christiane Fellbaum

Meaningful Texts
The Extraction of Semantic Information from Monolingual and Multilingual
Corpora
Edited by Geoff Barnbrook, Pernilla Danielsson and Michaela Mahlberg

Rethinking Idiomaticity
A Usage-based Approach
Stefanie Wulff

Working with Spanish Corpora
Edited by Giovanni Parodi

Studies in Corpus and Discourse

Corpus Linguistics and The Study of Literature
Stylistics In Jane Austen's Novels
Bettina Starcke

English Collocation Studies
The OSTI Report
John Sinclair, Susan Jones and Robert Daley
Edited by Ramesh Krishnamurthy
With an introduction by Wolfgang Teubert

Text, Discourse, and Corpora. Theory and Analysis
Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert
With an introduction by John Sinclair

This page intentionally left blank

Academic Vocabulary in
Learner Writing
From Extraction to Analysis

Magali Paquot

B L O O M S B U R Y
LONDON • NEW DELHI • NEW YORK • SYDNEY

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

50 Bedford Square
London
WC1B 3DP
UK

1385 Broadway
New York
NY 10018
USA

www.bloomsbury.com

Bloomsbury is a registered trade mark of Bloomsbury Publishing Plc

First published 2010 by Continuum
Paperback edition first published 2012

© Magali Paquot 2010

Magali Paquot has asserted her right under the Copyright, Designs and Patents Act, 1988, to be identified as Author of this work.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

No responsibility for loss caused to any individual or organization acting on or refraining from action as a result of the material in this publication can be accepted by Bloomsbury or the author.

Visit www.bloomsbury.com to find out more about our authors and their books
You will find extracts, author interviews, author events and you can sign up for newsletters to be the first to hear about our latest releases and special offers.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN: HB: 978-1-4411-3036-5
PB: 978-1-4411-1450-1
ePDF: 978-1-4411-9114-4
ePUB: 978-1-4411-0219-5

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress.

Contents

<i>Acknowledgements</i>	xi
<i>List of abbreviations</i>	xiii
<i>List of figures</i>	xv
<i>List of tables</i>	xvii
Introduction	1
Part I: Academic vocabulary	
Chapter 1 What is academic vocabulary?	9
1.1. Academic vocabulary vs. core vocabulary and technical terms	9
1.1.1. Core vocabulary	10
1.1.2. Academic vocabulary	11
1.1.3. Technical terms	13
1.1.4. Fuzzy vocabulary categories	13
1.2. Academic vocabulary and sub-technical vocabulary	17
1.3. Vocabulary and the organization of academic texts	22
1.4. Is there an ‘academic vocabulary’?	25
1.5. Summary and conclusion	27
Chapter 2 A data-driven approach to the selection of academic vocabulary	29
2.1. Corpora of academic writing	31
2.2. Corpus annotation	34
2.2.1. Issues in annotating corpora	34
2.2.2. The software	36
2.3. Automatic extraction of potential academic words	44
2.3.1. Keyness	46
2.3.2. Range	48
2.3.3. Evenness of distribution	50
2.3.4. Broadening the scope of well-represented semantic categories	53
2.4. The <i>Academic Keyword List</i>	55
2.5. Summary and conclusion	61

Part II: Learners' use of academic vocabulary

Chapter 3 Investigating learner language	67
3.1. The <i>International Corpus of Learner English</i>	67
3.2. Contrastive Interlanguage Analysis	70
3.3. A comparison of learner vs. expert writing	72
3.4. Summary and conclusion	78
Chapter 4 Rhetorical functions in expert academic writing	81
4.1. The <i>Academic Keyword List</i> and rhetorical functions	81
4.2. The function of exemplification	88
4.2.1. Using prepositions, adverbs and adverbial phrases to exemplify	90
4.2.2. Using nouns and verbs to exemplify	95
4.2.3. Discussion	106
4.3. The phraseology of rhetorical functions in expert academic writing	108
4.4. Summary and conclusion	122
Chapter 5 Academic vocabulary in the <i>International Corpus of Learner English</i>	125
5.1. A bird's-eye view of exemplification in learner writing	125
5.2. Academic vocabulary and general interlanguage features	142
5.2.1. Limited lexical repertoire	142
5.2.2. Lack of register awareness	150
5.2.3. The phraseology of academic vocabulary in learner writing	154
5.2.4. Semantic misuse	168
5.2.5. Chains of connective devices	174
5.2.6. Sentence position	177
5.3. Transfer-related effects on French learners' use of academic vocabulary	181
5.4. Summary and conclusion	192

Part III: Pedagogical implications and conclusions

Chapter 6 Pedagogical implications	201
6.1. Teaching-induced factors	201
6.2. The role of the first language in EFL learning and teaching	203
6.3. The role of learner corpora in EAP materials design	206

Chapter 7 General Conclusion	211
7.1. Academic vocabulary: a chimera?	211
7.2. Learner corpora, interlanguage and second language acquisition	215
7.3. Avenues for future research	216
<i>Appendix 1: Expressing cause and effect</i>	219
<i>Appendix 2: Comparing and contrasting</i>	226
<i>Notes</i>	235
<i>References</i>	240
<i>Author index</i>	257
<i>Subject index</i>	261

This page intentionally left blank

Acknowledgements

There are several people without whom this book would never have been written. First and foremost, I want to express my deepest and most sincere gratitude to my PhD supervisor, Professor Sylviane Granger, for her infectious enthusiasm, her intellectual perceptiveness and her unfailing expert guidance. I am greatly indebted to you, Sylviane, for giving me the opportunity to join the renowned Centre for English Corpus Linguistics seven years ago now! I have been lucky enough to undertake research in an environment where writing a PhD also means collaborating with many fellow researchers on up-and-coming projects, attending thought-provoking conferences, organizing seminars, conferences and summer schools, as well as lecturing and offering guidance to undergraduate students.

I am also very grateful to my colleagues and friends at the Centre for English Corpus Linguistics - Céline, Claire, Fanny, Gaëtanelle, Jennifer, Marie-Aude, Suzanne and Sylvie – for making the Centre for English Corpus Linguistics such an inspiring and intellectually stimulating research centre. I also wish to thank them for their moral and intellectual support and for all the entertaining lunchtimes we spent together talking about everyday life . . . and work.

I am indebted to a great number of colleagues not only for supplying me with corpora, corpus-handling tools and references, but also for providing helpful comments on earlier versions and stimulating ideas for my research. I would like to thank Yves Bestgen, Liesbet Degand, Jean Heiderscheidt, Sebastian Hoffmann, Scott Jarvis, Jean-René Klein, Fanny Meunier, Hilary Nesi, John Osborne and JoAnne Neff van Aertselaer. I am also grateful to an anonymous reviewer for recommendations on the first draft of the text.

I gratefully acknowledge the support of both the Communauté française de Belgique, which funded my doctoral dissertation out of which this book has grown, and the Belgian National Fund for Scientific Research (F.N.R.S).

On a more personal note, I would like to express my deepest thanks to my parents and friends for everything they have done to help me while I was working on this book. And last, but not least, Arnaud: thank you for making it all worthwhile.

Magali Paquot
Louvain-la-Neuve
November, 2009

List of abbreviations

AKL	Academic Keyword List (my own list)
AWL	Academic Word List (Coxhead, 2000)
BAWE	British Academic Written English (BAWE) Pilot Corpus
BNC	British National Corpus
B-BNC	Baby BNC Academic Corpus
BNC-AC	British National Corpus – academic sub-corpus
BNC-AC-HUM	British National Corpus – academic sub-corpus (discipline: humanities and arts)
BNC-SP	British National Corpus – spoken sub-corpus
CALL	Computer-assisted language learning
CECL	Centre for English Corpus Linguistics, Université catholique de Louvain
CIA	Contrastive Interlanguage Analysis
CLAWS	Constituent Likelihood Automatic Word-tagging system
CODIF	Corpus de Dissertations Françaises
EAP	English for academic purposes
EFL	English as a foreign language
ESL	English as a second language
ESP	English for specific purposes
GSL	General Service List (West, 1953)
ICLE	International Corpus of Learner English (Granger et al., 2002)
ICLEv2	International Corpus of Learner English (version 2) (Granger et al., 2009)
IL	interlanguage
L1	First language
L2	Foreign language
LDOCE4	Longman Dictionary of Contemporary English (4th edition)

LOCNESS	Louvain Corpus of Native Speaker Essays
LogL	Log-likelihood statistical test
MC	Micro-Concord Corpus Collection B
MED2	Macmillan English Dictionary for Advanced Learners (second edition)
MLD	Monolingual learners' dictionary
NS	Native speaker
NNS	Non-native speaker
pmw	Per million words
POS	Part-of-speech
SLA	Second language acquisition
UCREL	University Centre for Computer Corpus Research on Language, Lancaster University
WST4	WordSmith Tools (version 4)

List of figures

Figure 1.1:	The relationship between academic and sub-technical vocabulary	21
Figure 2.1:	A three-layered sieve to extract potential academic words	45
Figure 2.2:	WordSmith Tools – WordList option	49
Figure 2.3:	Distribution of the words example and law in the 15 sub-corpora	50
Figure 2.4:	WordSmith Tools Detailed Consistency Analysis	51
Figure 2.5:	Distribution of the noun ‘solution’	53
Figure 3.1:	ICLE task and learner variables (Granger et al., 2002: 13)	68
Figure 3.2:	Contrastive Interlanguage Analysis (Granger 1996a)	70
Figure 3.3:	BNCweb Collocations option	77
Figure 4.1:	Exemplification in the BNC-AC-HUM	89
Figure 4.2:	The distribution of the adverb ‘notably’ across genres	93
Figure 4.3:	The distribution of ‘by way of illustration’ across genres	94
Figure 4.4:	The distribution of ‘to name but a few’ across genres	95
Figure 4.5:	The distribution of the verbs ‘illustrate’ and ‘exemplify’ across genres	103
Figure 4.6:	The phraseology of rhetorical functions in academic prose	121
Figure 5.1:	Exemplifiers in the ICLE and the BNC-AC-HUM	127
Figure 5.2:	The use of the prepositions ‘like’ and ‘such as’ in different genres	131
Figure 5.3:	The use of the adverb ‘notably’ in different genres	131

Figure 5.4:	Distribution of the adverbials ‘for example’ and ‘for instance’ across genres in the BNC	132
Figure 5.5:	The treatment of ‘namely’ on websites devoted to English connectors	140
Figure 5.6:	The use of ‘despite’ and ‘in spite of’ in different genres	145
Figure 5.7:	The frequency of speech-like lexical items in expert academic writing, learner writing and speech (based on Gilquin and Paquot, 2008)	153
Figure 5.8:	Phraseological cascades with ‘in conclusion’ and learner-specific equivalent sequences	161
Figure 5.9:	Collocational overlap	165
Figure 5.10:	A possible rationale for the use of ‘according to me’ in French learners’ interlanguage	187
Figure 5.11:	A possible rationale for the use of ‘let us in French learners’ interlanguage	191
Figure 5.12:	Features of novice writing - Frequency in expert academic writing, native-speaker and EFL novices’ writing and native speech (per million words of running text)	195
Figure 6.1:	Connectives: contrast and concession (Jordan 1999:136)	202
Figure 6.2:	Comparing and contrasting: using nouns such as ‘resemblance’ and ‘similarity’ (Gilquin et al., 2007b: IW5)	208
Figure 6.3:	Reformulation: Explaining and defining: using ‘i.e.’, ‘that is’ and ‘that is to say’ (Gilquin et al., 2007b: IW9)	209
Figure 6.4:	Expressing cause and effect: ‘Be careful’ note on ‘so’ (Gilquin et al., 2007b: IW13)	210

List of tables

Table 1.1:	Composition of the Academic Corpus (Coxhead 2000: 220)	12
Table 1.2:	Chung and Nation's (2003: 105) rating scale for finding technical terms, as applied to the field of anatomy	14
Table 1.3:	Word families in the AWL	17
Table 2.1:	The corpora of professional academic writing	31
Table 2.2:	The re-categorization of data from the professional corpus into knowledge domains	32
Table 2.3:	The corpora of student academic writing	33
Table 2.4:	Examples of essay topics in the BAWE pilot corpus	34
Table 2.5:	An example of CLAWS vertical output	39
Table 2.6:	CLAWS horizontal output [lemma + POS]	40
Table 2.7:	CLAWS horizontal output [lemma + simplified POS tags]	40
Table 2.8:	Simplification of CLAWS POS-tags	41
Table 2.9:	CLAWS tagging of the complex preposition 'in terms of'	41
Table 2.10:	Semantic fields of the UCREL Semantic Analysis System	42
Table 2.11:	USAS vertical output	43
Table 2.12:	USAS horizontal output	44
Table 2.13:	The fiction corpus	47
Table 2.14:	Number of keywords	47
Table 2.15:	Automatic semantic analysis of potential academic words	54
Table 2.16:	Distribution of grammatical categories in the <i>Academic Keyword List</i>	55
Table 2.17:	The <i>Academic Keyword List</i>	56
Table 2.18:	The distribution of AKL words in the GSL and the AWL	60

Table 3.1:	Breakdown of ICLE essays	69
Table 3.2:	BNC Index – Breakdown of written BNC genres (Lee 2001)	74
Table 4.1:	Ways of expressing exemplification found in the BNC-AC-HUM	89
Table 4.2:	The use of ‘for example’ and ‘for instance’ in the BNC-AC-HUM	91
Table 4.3:	The use of ‘example’ and ‘for example’ in the BNC-AC-HUM	95
Table 4.4:	Significant verb co-occurents of the noun ‘example’ in the BNC-AC-HUM	96
Table 4.5:	Adjective co-occurents of the noun ‘example’ in the BNC-AC-HUM	100
Table 4.6:	The use of the lemma ‘illustrate’ in the BNC-AC-HUM	103
Table 4.7:	The use of the lemma ‘exemplify’ in the BNC-AC-HUM	105
Table 4.8:	The use of imperatives in academic writing (based on Siepmann, 2005: 119)	107
Table 4.9:	Ways of expressing a concession in the BNC-AC-HUM	109
Table 4.10:	Ways of reformulating, paraphrasing and clarifying in the BNC-AC-HUM	109
Table 4.11:	Ways of expressing cause and effect in the BNC-AC-HUM	110
Table 4.12:	Ways of comparing and contrasting found in the BNC-AC-HUM	112
Table 4.13:	Co-occurents of nouns expressing cause or effect in the BNC-AC-HUM	115
Table 4.13a:	<i>reason</i>	115
Table 4.13b:	<i>implication</i>	115
Table 4.13c:	<i>effect</i>	116
Table 4.13d:	<i>outcome</i>	116
Table 4.13e:	<i>result</i>	117
Table 4.13f:	<i>consequence</i>	117
Table 4.14:	Co-occurents of verbs expressing possibility and certainty in the BNC-AC-HUM	119
Table 4.14a:	suggest	119
Table 4.14b:	prove	120
Table 4.14c:	appear	120
Table 4.14d:	tend	120

Table 5.1:	A comparison of exemplifiers based on the total number of running words	128
Table 5.2:	A comparison of exemplifiers based on the total number of exemplifiers used	129
Table 5.3:	Two methods of comparing the use of exemplifiers	130
Table 5.4:	Significant adjective co-occurrences of the noun 'example' in the ICLE	133
Table 5.5:	Adjectives co-occurrences of the noun 'example' in ICLE not found in the BNC	133
Table 5.6:	Significant verb co-occurrences of the noun 'example' in the ICLE	134
Table 5.7:	Verb co-occurrence types of the noun 'example' in ICLE not found in BNC	134
Table 5.8:	The distribution of 'example' and 'be' in the ICLE and the BNC-AC-HUM	135
Table 5.9:	The distribution of 'there + BE + example' in ICLE and the BNC-AC-HUM	135
Table 5.10:	The distribution of AKL words in the ICLE	143
Table 5.11:	Examples of AKL words which are overused and underused in the ICLE	144
Table 5.12:	Two ways of comparing the use of cause and effect markers in the ICLE and the BNC	146
Table 5.13:	The over- and underuse by EFL learners of specific devices to express cause and effect (based on Appendix 1)	147
Table 5.14:	The over- and underuse by EFL learners of specific devices to express comparison and contrast (based on Appendix 2)	149
Table 5.15:	Speech-like overused lexical items per rhetorical function	151
Table 5.16:	The frequency of 'maybe' in learner corpora	154
Table 5.17:	The frequency of 'I think' in learner corpora	154
Table 5.18:	Examples of overused and underused clusters with AKL words	156
Table 5.19:	Clusters of words including AKL verbs which are over- and underused in learners' writing, by comparison with expert academic writing	158
Table 5.20:	Examples of overused clusters in learner writing	159
Table 5.21:	Verb co-occurrences of the noun <i>conclusion</i> in the ICLE	162

Table 5.22:	Adjective co-occurents of the noun <i>conclusion</i> in the ICLE	167
Table 5.23:	The frequency of sentence-initial position of connectors in the BNC-AC-HUM and the ICLE	178
Table 5.24:	Sentence-final position of connectors in the ICLE and the BNC-AC-HUM	181
Table 5.25:	Jarvis's (2000) three effects of potential L1 influence	183
Table 5.26:	Jarvis's (2000) unified framework applied to the ICLE-FR	184
Table 5.27:	A comparison of the use of the English verb 'illustrate' and the French verb 'illustrer'	188
Table 5.28:	'let us' in learner texts	189
Table 5.29:	The transfer of frequency of the first person plural imperative between French and English writing	191
Table 6.1:	Le Robert & Collins CD-Rom (2003–2004): Essay writing	205

Introduction

That English has become the major international language for research and publication is beyond dispute. As a result, university students need to have good receptive command of English if they want to have access to the literature pertaining to their discipline. As a large number of them are also required to write academic texts (e.g. essays, reports, MA dissertations, PhD theses, etc.), they also need to have a productive knowledge of academic language. As noted by Biber, 'students who are beginning university studies face a bewildering range of obstacles and adjustments, and many of these difficulties involve learning to use language in new ways' (2006: 1). Several studies have shown that the distinctive, highly routinized, nature of academic prose is problematic for many novice native-speaker writers (e.g. Cortes, 2002), but poses an even greater challenge to students for whom English is a second (e.g. Hinkel, 2002) or foreign language (e.g. Gilquin et al., 2007b).

Studies in second language writing have established that learning to write second-language (L2) academic prose requires an advanced linguistic competence, without which learners simply do not have the range of lexical and grammatical skills required for academic writing (Jordan, 1997; Nation and Waring, 1997; Hinkel, 2002; 2004; Reynolds, 2005). A questionnaire survey of almost 5,000 undergraduates showed that students from all 26 departments at the Hong Kong Polytechnic University experienced difficulties with the writing skills necessary for studying content subjects through the medium of English (Evans and Green, 2006). Almost 50 per cent of the students reported that they encountered difficulties in using appropriate academic style, expressing ideas in correct English and linking sentences smoothly. Mastering the subtleties of academic prose is, however, not only a problem for novice writers. International refereed journal articles are regarded as the most important vehicle for publishing research findings and non-native academics who want to publish their work in those top journals often find their articles rejected, partly because of language problems.

These problems include the fact that they have less facility of expression and a poorer vocabulary; they find it difficult to 'hedge' appropriately and the structure of their texts may be influenced by their first language (see Flowerdew, 1999).

Because it causes major difficulties to students and scholars alike, academic discourse has become a major object of study in applied linguistics. Flowerdew (2002) identified four major research paradigms for investigating academic discourse, namely (Swalesian) genre analysis, contrastive rhetoric, ethnographic approaches and corpus-based analysis. While the first three approaches to English for Academic Purposes (EAP) emphasize the situational or cultural context of academic discourse, corpus-linguistic methods focus more on the co-text of selected lexical items in academic texts.

Corpus linguistics is concerned with the collection in electronic format and the analysis of large amounts of naturally occurring spoken or written data 'selected according to external criteria to represent, as far as possible, a language or language variety as a source of linguistic research' (Sinclair, 2005: 16). Computer corpora are analysed with the help of software packages such as WordSmith Tools 4 (Scott, 2004), which includes a number of text-handling tools to support quantitative and qualitative textual data analysis. Wordlists give information on the frequency and distribution of the vocabulary – single words but also word sequences – used in one or more corpora. Wordlists for two corpora can be compared automatically so as to highlight the vocabulary that is particularly salient in a given corpus, i.e., its keywords. Concordances are used to analyse the co-text of a linguistic feature, in other words its linguistic environment in terms of preferred co-occurrences and grammatical structures. The research paradigm of corpus linguistics is ideally suited for studying the linguistic features of academic discourse as it can highlight which words, phrases or structures are most typical of the genre and how they are generally used.

Corpus-based studies have already shed light on a number of distinctive linguistic features of academic discourse as compared with other genres. Biber's (1988) study of variation across speech and writing has shown that academic texts typically have an informational and non-narrative focus; they require highly explicit, text-internal reference and deal with abstract, conceptual or technical subject matter (Biber, 1988: 121–60). The *Longman Grammar of Spoken and Written English* (Biber et al., 1999) provides a comprehensive description of the range of distinctive grammatical and lexical features of academic prose, compared to conversation, fiction and newspaper reportage. Common features of this genre include a high rate of

occurrence of nouns, nominalizations, noun phrases with modifiers, attributive adjectives, derived adjectives, activity verbs, verbs with inanimate subjects, agentless passive structures and linking adverbials. By contrast, first and second person pronouns, private verbs, *that*-deletions and contractions occur very rarely in academic texts.

In addition, studies of vocabulary have emphasized the importance of a 'sub-technical' or 'academic' vocabulary alongside core words and technical terms in academic discourse (Nation, 2001: 187–216). Hinkel (2002: 257–65) argues that the exclusive use of a process-writing approach, the relative absence of direct and focused grammar instruction, and the lack of academic vocabulary development contribute to a situation in which non-native students are simply not prepared to write academic texts. She provides a list of priorities in curriculum design and writes that, among the top priorities, 'NNSs [non-native students] need to learn more contextualized and advanced academic vocabulary, as well as idioms and collocations to develop a substantial lexical arsenal to improve their writing in English' (Hinkel, 2002: 247). The *Academic Word List* (Coxhead, 2000) was compiled on the basis of corpus data to meet the specific vocabulary needs of students in higher education settings.

But what is 'academic vocabulary'? Despite its widespread use, the term has been used in various ways to refer to different (but often overlapping) vocabulary categories. This book aims to provide a better description of the notion of 'academic vocabulary'. It takes the reader full circle, from the extraction of potential academic words through their linguistic analysis in expert and learner corpus data, to the pedagogical implications that can be drawn from the results. Recent corpus-based studies have emphasized the specificity of different academic disciplines and genres. As a result, researchers such as Hyland and Tse (2007) question the widely held assumption that students need a common core vocabulary for academic study. They argue that the different disciplinary literacies undermine the usefulness of such lists and recommend that lecturers help students develop a discipline-based lexical repertoire.

This book is an attempt to resolve the tension between the particularizing trend which advocates the teaching of a more restricted, discipline-based vocabulary syllabus, and the generalizing trend which recognizes the existence of a common core 'academic vocabulary' that can be taught to a large number of learners in many disciplines. I first argue that, to resolve this tension, the concept of 'academic vocabulary' must be revisited. I demonstrate, on the basis of corpus data, that, as well as discipline-specific vocabulary, there is a wide range of words and phraseological patterns that

are used to refer to activities which are characteristic of academic discourse, and more generally, of scientific knowledge, or to perform important discourse-organizing or rhetorical functions in academic writing.

A large proportion of this lexical repertoire consists of core vocabulary, a category which has so far been largely neglected in EAP courses but which is usually not fully mastered by English as a foreign language (EFL) learners, even those at the high-intermediate or advanced levels. I make use of Granger's (1996a) *Contrastive Interlanguage Analysis* to test the working hypothesis that upper-intermediate to advanced EFL learners, irrespective of their mother tongue background, share a number of linguistic features that characterize their use of academic vocabulary. The learner corpus used is the first edition of the *International Corpus of Learner English* (ICLE), which is among the largest non-commercial learner corpora in existence. It contains texts written by learners with different mother tongue backgrounds. Ten ICLE sub-corpora representing different mother tongue backgrounds (Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish) are compared with a subset of the academic component of the *British National Corpus* (texts written by specialists in the Humanities) to identify ways in which learners' use of academic vocabulary differs from that of more expert writers. A comparison of the ten sub-corpora then makes it possible to identify linguistic features that are shared by learners from a wide range of mother tongue backgrounds, and therefore possibly developmental. The EFL learners are all learning how to write in a foreign language, and they are often novice writers in their mother tongue as well.

However, not all learner specific-features can be attributed to developmental factors. The comparison of several ICLE sub-corpora helps to pinpoint a number of patterns that are characteristic of learners who share the same first language, and which may therefore be transfer-related. I made use of Jarvis's (2000) unified framework to investigate the potential influence of the first language on French learners' use of academic vocabulary in English.

The book is organized in three sections. The first scrutinizes the concept of 'academic vocabulary', reviewing the many definitions of the term and arguing that, for productive purposes, academic vocabulary is more usefully defined as a set of options to refer to those activities that characterize academic work, organize scientific discourse, and build the rhetoric of academic texts. It then proposes a data-driven procedure based on the criteria of keyness, range, and evenness of distribution, to select academic words that could be part of a common core academic vocabulary syllabus.

The resulting list, called the *Academic Keyword List (AKL)*, comprises a set of 930 potential academic words. One important feature of the methodology is that, unlike Coxhead's (2000) *Academic Word List*, the AKL includes the 2,000 most frequent words of English, thus making it possible to appreciate the paramount importance of core English words in academic prose.

The AKL is used in Section 2 to explore the importance of academic vocabulary in expert writing and to analyse EFL learners' use of lexical devices that perform rhetorical or organizational functions in academic writing. This section offers a thorough analysis of these lexical devices as they appear in the *International Corpus of Learner English*, describing the factors that account for learners' difficulties in academic writing. These factors include a limited lexical repertoire, lack of register awareness, infelicitous word combinations, semantic misuse, sentence-initial positioning of adverbs and transfer effects.

The final section briefly comments on the pedagogical implications of these results, summarizes the major findings, and points the way forward to further research in the area.

This page intentionally left blank

Part I

Academic vocabulary

‘Academic vocabulary’ is a term that is widely used in textbooks on English for academic purposes and Second Language Acquisition (SLA) reference books. Nevertheless, it can be understood in a variety of ways and used to indicate different categories of vocabulary. In this section, my objectives are to clarify the meaning of ‘academic vocabulary’ by critically examining its many uses, and to build a list of words that fit my own definition of the term. Chapter 1 therefore tries to identify the key features of academic vocabulary and to clear up the confusion between academic words and other vocabulary. Chapter 2 proposes a data-driven methodology based on the criteria of keyness, range and evenness of distribution, and uses this to build a new list of potential academic words, viz. the *Academic Keyword List* (AKL). This list is very different from Coxhead’s *Academic Word List* and has already been used to inform the writing sections in the second edition of the *Macmillan English Dictionary for Advanced Learners* (see Gilquin et al., 2007b). The AKL is used in Section 2 to analyse EFL learners’ use of lexical devices that perform rhetorical or organizational functions in academic writing.

This page intentionally left blank