

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

From Corpus to Classroom:
language use and
language teaching

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

From Corpus to Classroom: language use and language teaching

ANNE O’KEEFFE
MICHAEL MCCARTHY
& RONALD CARTER



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

www.cambridge.org

Information on this title: www.cambridge.org/9780521616867

© Cambridge University Press 2007

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2007

Reprinted 2010

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-85146-6 Hardback

ISBN 978-0-521-61686-7 Paperback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to in
this publication, and does not guarantee that any content on such websites is,
or will remain, accurate or appropriate. Information regarding prices, travel
timetables and other factual information given in this work are correct at
the time of first printing but Cambridge University Press does not guarantee
the accuracy of such information thereafter.

Acknowledgements

In writing this book, we have received help, support and inspiration from many sources. First and foremost, we thank Alison Sharpe, Associate Publishing Director, ELT, at Cambridge University Press, who decided to run with this particular idea and who has been a constant source of support to our work over many years. We are also extremely grateful to Jane Walsh, Senior Development Editor, for managing this endeavour. We appreciated greatly her support from start to finish. Thanks also to Geraldine Mark who edited the book and took it through its final stages.

This book stands on the shoulders of a huge amount of work over the last thirty years in the areas of corpus linguistics and applied linguistics. Developments in corpus linguistics have inspired each of us in how we look at language, how we design materials and how we teach, and research in applied linguistics has offered us broad frameworks in which to make sense of it all. We therefore acknowledge the work that has been done to bring us to where we are. Above all, we acknowledge the work of John Sinclair. Every chapter of this book is influenced by his ideas. For each of us, he has generously inspired and nurtured our work over the years. The work of Luke Prodromou is also very influential for us in this book. His work on ‘Successful Users of English’ provides a paradigm shift in how we view English language use in a global context, and one which is particularly salient in current debates.

This book is the first to come out of the Inter-Varietal Applied Corpus Studies (IVACS) inter-institutional collaboration between the University of Limerick, Ireland, the University of Nottingham and the Queen’s University Belfast, UK. What brings us together is reflected in this book: an interest in the applications of corpus linguistics for the analysis of language in use and what this can tell us about how and what we teach. We acknowledge our colleagues in IVACS for their part in making this book happen: Svenja Adolphs, Carolina Amador Moreno, James Binchy, Brian Clancy, Jane Evison, Fiona Farr, Loretta Fung, Michael Handford, Dawn Knight, Barbara Malveira Orfanó, Bróna Murphy, Róisín Ní Mhócháin, Aisling O’Boyle, María Palma Fahey, Nikoleta Rapti, Paul Roberts, Norbert Schmitt, Ivor Timmis, Elaine Vaughan, Steve Walsh and Wang Shih-Ping. Other colleagues and friends who have inspired us during the course of writing this book include Angela Chambers, Winnie Cheng, Paul Heacock, Michael Hoey, Almut Koester, James Lantolf, Nigel McQuitty, Rosamund Moon, Jeanne McCarten, Felicity O’Dell, Barry O’Sullivan, Randi Reppen, Helen Sandiford, Howard Siegelman, Peter Stockwell, Steve Thorne, Koen Van Landeghem, Mary Vaughn and Martin Warren.

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

vi From Corpus to Classroom: language use and language teaching

We owe a huge debt of gratitude to Susan Hunston, who provided detailed comments and constructive criticism for us on the first draft of the manuscript. The final version of this book has benefited enormously from her clear and generous feedback. We are also grateful to Dave Evans for his extensive work with us on the index. As the cliché goes, responsibility for any inadequacies which remain in the book rests firmly at the door of the authors.

Most of all we thank our respective partners, Ger Downes, Jeanne McCarten and Jane Carter, without whose support this book would have no meaning.

ANNE O’KEEFFE
MICHAEL MCCARTHY
RONALD CARTER

Cover design: David Lawton. Text design: Dale Tomlinson.

Contents

Acknowledgements v

Preface xi

1 Introduction 1

- 1.1 Introduction: the basics 1
- 1.2 What is a corpus and how can we use it? 1
- 1.3 Which corpus, what for and what size? 3
- 1.4 How to make a basic corpus 5
- 1.5 Basic corpus linguistic techniques 8
- 1.6 Lexico-grammatical profiles 14
- 1.7 How have corpora been used? 17
- 1.8 How have corpora influenced language teaching? 21
- 1.9 Issues and debates in the use of corpora in language teaching 25

2 Establishing basic and advanced levels in vocabulary learning 31

- 2.1 Introduction 31
- 2.2 Frequency and native-speaker vocabulary size 31
- 2.3 The most frequent words and the core vocabulary 33
- 2.4 The broad categories of a basic vocabulary 37
- 2.5 Chunks at the basic level 46
- 2.6 The basic level: conclusion 46
- 2.7 The advanced level 47
- 2.8 Targets 48
- 2.9 The vocabulary curve 49
- 2.10 The 6,000 to 10,000 word band 50
- 2.11 Meanings and connotations 53
- 2.12 Breadth and depth 54

3 Lessons from the analysis of chunks 58

- 3.1 Introduction 58
- 3.2 The single word 58
- 3.3 Collocation 59
- 3.4 Strings of words in corpora 60
- 3.5 Phraseology and idiomaticity 62
- 3.6 Looking at corpus data 64
- 3.7 Interpreting the data: chunks and single words 69
- 3.8 Chunks and units of interaction 70
- 3.9 Conclusions and implications 75

viii From Corpus to Classroom: language use and language teaching

- 4 Idioms in everyday use and in language teaching 80**
- 4.1 Introduction 80
- 4.2 Finding and classifying idioms 82
- 4.3 Frequency 84
- 4.4 Meaning 86
- 4.5 Functions of idioms 87
- 4.6 Idioms in specialised contexts 90
- 4.7 Idioms in teaching and learning 94
- 5 Grammar and lexis and patterns 100**
- 5.1 Introduction 100
- 5.2 The example of *border* 102
- 5.3 Grammar rules and patterns: deterministic and probabilistic 104
- 5.4 The *get*-passive: an extended case study 106
- 5.5 Previous studies of the *get*-passive 106
- 5.6 *Get*-passives and related forms 108
- 5.7 Core *get*-passive constructions in the CANCODE sub-corpus 109
- 5.8 Discussion 113
- 5.9 Grammar as structure and grammar as probabilities: the example of ellipsis 114
- 5.10 Conclusions and implications 115
- 6 Grammar, discourse and pragmatics 120**
- 6.1 Introduction 120
- 6.2 Non-restrictive *which*-clauses 120
- 6.3 Previous studies of *which*-clauses 122
- 6.4 Concordance analysis of *which*-clauses 122
- 6.5 *If*-clauses 127
- 6.6 *Wh*-cleft clauses 130
- 6.7 Bringing the insights together 136
- 6.8 Corpus grammar and pedagogy 137
- 7 Listenership and response 140**
- 7.1 Introduction 140
- 7.2 Forms of listenership 142
- 7.3 Response tokens across varieties of English 145
- 7.4 Functions of response tokens 148
- 7.5 Conclusions and implications 155
- 8 Relational language 159**
- 8.1 Introduction 159
- 8.2 Conversational routines 163
- 8.3 Small talk 168
- 8.4 Discourse markers 171
- 8.5 Hedging 174

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

8.6	Vagueness and approximation	176
8.7	Conclusions and implications	181
9	Language and creativity: creating relationships	184
9.1	Introduction	184
9.2	Spoken language and creativity	184
9.3	Corpora and creativity	188
9.4	Creative speakers	190
9.5	Applications to pedagogy	191
9.6	Corpus to pedagogy: creating relationships	192
9.7	SUEs and creativity	192
9.8	Quantitative and qualitative	196
9.9	Conclusions	197
10	Specialising: academic and business corpora	198
10.1	Introduction	198
10.2	Written academic English	198
10.3	Written academic English: examples of frequency	200
10.4	Spoken academic corpora	203
10.5	Spoken academic English, conversation and spoken business English	204
10.6	The CANBEC business corpus	206
10.7	Chunks	210
10.8	<i>Problem</i> and its institutional construction in CANBEC	214
10.9	Summary	216
10.10	Pedagogical implications	216
11	Exploring teacher corpora	220
11.1	Introduction	220
11.2	Classroom discourse	222
11.3	Frameworks for the analysis of classroom language	222
11.4	Applying the frameworks to a corpus of classroom data	229
11.5	Looking at questioning in the classroom	233
11.6	Teacher corpora in professional development	240
11.7	Conclusions and considerations	243
	<i>Coda</i>	246
	<i>References</i>	249
	<i>Appendix 1</i>	284
	<i>Appendix 2</i>	297
	<i>Appendix 3</i>	301
	<i>Author index</i>	305
	<i>Subject index</i>	310
	<i>Publisher’s acknowledgements</i>	314

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

Preface

In recent years, conferences on applied linguistics and teacher development, as well as published material such as books, articles and newsletters, frequently refer to developments and findings in the field of corpus linguistics. An increasing number of materials and resources for use in language teaching and learning now boast that they are ‘corpus-based’ or ‘corpus-informed’. Indeed, in the pioneering area of learners’ dictionaries, one could hardly imagine any major publisher nowadays putting out a dictionary that was not based on a corpus, such was the revolution sparked off by Sinclair’s COBUILD dictionary project in the 1980s. Similarly, corpus information, in recent years, seems to be becoming *de rigueur* as the basis of the compilation of major reference grammars, and, more and more, as a major feature of coursebooks, though here the picture is more patchy at the time of writing.

However, widespread use of ‘corpus linguistics’ does not mean that the term or its findings are necessarily fully or widely understood in the context of language pedagogy. In addition, many important developments in the field of corpus linguistics are not always communicated or usefully mediated in terms of their implications for language teaching. This is possibly because corpus linguists are very often not language teachers and spend a lot of time talking with one another rather than with teachers. This book aims to address the frequent mismatch between corpus linguistics research and what goes into materials and resources, and what goes on in the language classroom. It aims to highlight the outcomes which we consider to be relevant and transferable in terms of how they can inform pedagogy, or challenge how and what we teach. But the book stops at the classroom door. We do not intend to tell you how to teach and what to do in your own classes; only you can know best what is effective and appropriate in your specific local context, and you are by far the best person to take the final, practical steps in applying our ‘applied’ linguistics, if you judge the book to have value.

Not all descriptive findings about language are of relevance to how and what we teach, but very many of them are. Here we aim to start with the basics. We do not assume any prior knowledge or experience of corpus linguistics. The book begins by explaining what is meant by a corpus, how one is made, and the most common techniques that can be used to analyse language in a corpus. We also aim to identify what we see as key findings that may lead to new pedagogical insights for language teachers. In so doing, the book aims to provide the critical knowledge and stimulus for language teachers to get involved in the exciting area of corpus linguistics and to make informed decisions about corpus findings in terms of how, or whether, these can inform their teaching, translate into classroom practice, or inform

their choices of materials and other resources. Nowadays, given the bewildering range of available materials and the inevitable claims of publishers that theirs are the best, it helps more than ever to be able, calmly and confidently, to question and evaluate claims made about materials, especially in the relatively new area of corpus-informed ones.

We are aware that a book entitled *From Corpus to Classroom* promises many things. It is helpful, at this stage, to make clear what it is not. This book is not about data-driven learning (often referred to as DDL), that is, where data from language corpora (most typically concordances) are used in a hands-on manner in the classroom by the learners. There are many existing publications which address and facilitate this approach. This book is not about telling language teachers how to teach. We are not saying 'this is what it says in a corpus and so you have to teach it'. This book does not provide 'off-the-shelf' solutions or materials that can be rolled out in any and every classroom. It is about informing the reader of the relevant research that is on-going in the field of corpus linguistics and summarising the findings in terms of what we, its authors, consider to have relevance to language teaching. It is about making such research accessible by explaining key concepts, beginning with the assumption of zero background knowledge in the area. Our aim is to facilitate a discerning understanding of what it actually means when claims are made that such things as syllabuses, reference resources and teaching materials are 'corpus-based'.

Most of the chapters in this book draw primarily on spoken language corpora, so much so that at one point, we debated whether the word 'spoken' should be included in the title. However, given that most books on corpora draw primarily on written data and do not feel any need to make this explicit in their titles, we have decided not to apologise for our attempt to redress the balance. Most of our research, over the years, has endeavoured to challenge the dominance of the written word. We hope that this is also the case here. We are also very conscious in this book that there is a proliferation of corpora dedicated to the English language. Where possible we try to use as many types of Englishes as we have been able to access, and we sometimes refer to research that relates to languages other than English. We accept that we come nowhere near finding a balance, and could hardly do so in a book aimed at a wide international readership for whom English is typically the professional lingua franca, but we think that it is important to highlight this point at the outset. At the time of writing, there is far more corpus-based research into English than into any other language (see Wilson, Rayson and McEnery 2003 for more on corpora of languages other than English). Perhaps some of the readers of this book can contribute to redressing the imbalance by building on the existing work using non-English data.

The book opens with a foundational chapter which aims to provide the critical knowledge for building and using a corpus. It also focuses on key issues and debates that have emerged around corpus research. We feel these need to be addressed as a backdrop to the chapters which follow. These issues centre mostly around debates about authenticity and native speakers versus non-native speakers. We are conscious throughout the book to avoid absolutism in relation to native versus non-native speakers of a language. We take the position that the concept of the ideal native speaker is an ephemeral one, and we search in vain for that elusive phantom in our corpora. Real speakers whose utterances we analyse in

corpus examples are very often struggling with the demands of real-time communication. Indeed, if we compare the everyday human activities of talking and walking, talking has been compared to a series of uncertain lurches rather than to smooth walking (Krauss et al. 1995). We therefore find the term 'Successful User of English' (SUE), after the work of Luke Prodromou (2003a), to be a much more appropriate term than 'native speaker'. This is discussed and exemplified in chapter one.

All three authors of this book have been inspired by the seminal work of John Sinclair in the field of corpus linguistics, and the structure of the book is motivated by the importance that his work places on the word as the starting point for the description of meaning. As he puts it, 'the word is the unit that aligns grammar and vocabulary' (Sinclair 1996a: 75). Hence the body of the book is structured so that it moves from the word to everyday strings of words (or *chunks*) and idioms, then onto grammar, which subsequently leads us into pragmatics, discourse and creativity. Finally, the closing chapters of the book look at specialised corpora in the areas of teacher development and the institutional contexts of academic and business communication.

Chapter 2 looks at the most frequently occurring words in written and spoken English. It focuses on the pedagogical relevance of corpus findings in terms of our understanding of the vocabulary needs of second language learners. We explore how this information can be beneficial for establishing benchmarks by which learners' vocabulary levels can be assessed and by which we may come to some general agreement as to what constitutes the various levels of proficiency in vocabulary knowledge.

Chapter 3 brings us from the single word to clusters of words, or chunks. Corpus software can tell us what the most frequent chunks in a language are, but this information in its raw form is not terribly illuminating. This chapter proposes a functional categorisation for the most frequent items and explores some of the issues connected with working with chunks in the classroom.

Chapter 4 addresses idioms. This chapter gives consideration to how we define idioms and how they can be extracted from a corpus. This is a qualitative and interpretive process (a computer does not know what an idiom is), and one which we hope can be replicated by those interested in exploring this area further. We take a broad view of idioms and we believe the classification has transfer for the classroom and, particularly, for the design of materials for the teaching of idioms.

In the progression from the single word and lexical chunks, chapter 5 brings us to the next level 'up', that is the interface between lexis and grammar, or 'lexico-grammar'. The phraseological or lexico-grammatical patterns that we explore here, such as choices between *he's not* and *he isn't*, are found to be systematic and go beyond a straightforward grammatical description.

Chapter 6 brings us from phrasal- and clausal-level considerations to discourse and pragmatics. This is contextualised using two structures which are very familiar to language teachers: non-restrictive (sometimes called non-defining) *which*-clauses, and *if*-clauses. We aim to show how a corpus can reveal a lot about the pragmatic force of grammatical choices.

In chapter 7 we focus on one aspect of discourse which we see as having great relevance to language pedagogy and the promotion of fluency. Here we concentrate on the notion of *listenership*, whereby interaction is seen as a two way speaker-hearer process. For spoken discourse to be successful, it demands that the listener responds appropriately to the ongoing speaker turns. The markers of successful listenership are explored, using corpus data, both in terms of the typical structures that are used by listeners and in terms of how they can perform different functions.

Chapter 8 brings together all the chapters that precede it by focusing on how words, chunks and lexico-grammatical patterns can have relational functions. It focuses on areas of spoken language which, in the past, have mostly been the domain of pragmatics and conversation analysis, but which can be explored very effectively in both a quantitative and qualitative way using corpora (for example, small talk, conversational routines, hedging, vague language).

Chapter 9 explores corpus examples in terms of the everyday creativity of users and addresses how this can be appreciated and enjoyed in the classroom. This chapter is a good example of our attempts to redress the balance between spoken and written English. We are very used to talking about creativity in written prose and poetry, but rarely consider it in spoken language. Now that the ephemerality of the spoken word can be overcome by looking at spoken corpus data, we see this as an important contribution to the building of frameworks for looking at spoken language in this way. We also hope that this chapter will go some way to redress the bias towards the rather utilitarian views of language immanent in many versions of communicative language teaching.

Chapter 10 deals with academic and business corpora and what lessons these have for the courses that we teach and the materials that we use. Here both written and spoken data are used and high frequency vocabulary items are discussed. The chapter aims to show the value of smaller and specialised corpora in contrast to the ever-bigger, billion-word-plus corpora built by major publishers primarily to serve the needs of lexicographers.

The final chapter in the body of the book is intended to facilitate the use of corpora in teacher education and development. It is a very broad chapter in a number of ways, and indeed it differs from all the previous chapters. It is broad in the sense that it offers the possibility of a corpus as a collection of transcribed classroom interactions, even if it is just following one class or group of students. This is sufficient, we believe, as a starting point to using a corpus for teacher reflection. As little as one class can provide enough material to facilitate scrutiny of the commonest processes of classroom interaction. It is also broad in the sense that it provides three frameworks which can be used by teachers as the basis for reflecting on practice. None of these frameworks comes from corpus linguistics (and many of our readers may already be aware of them), but they all have much to offer to the interpretation of classroom discourse in a corpus. We end the book with a coda, which looks forward to the future.

We have enjoyed writing this book very much. It has challenged us to look at what we do and articulate its relevance and implications for pedagogy. We hope that by the end of the book you are as excited about what corpus linguistics has to offer language pedagogy as

Cambridge University Press

978-0-521-61686-7 - From Corpus to Classroom: Language use and Language Teaching

Anne O’Keeffe, Michael McCarthy and Ronald Carter

Frontmatter

[More information](#)

we are, and that the book will have bridged a conceptual gap, and facilitated access to an area of immense potential for language teachers, syllabus designers and materials writers and researchers in the area of applied linguistics.

ANNE O’KEEFFE
MICHAEL MCCARTHY
RONALD CARTER