Cambridge Textbooks in Linguistics

Corpus Linguistics

Tony McEnery and Andrew Hardie

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9780521838511

Corpus Linguistics

Corpus linguistics is the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts. This textbook outlines the basic methods of corpus linguistics, explains how the discipline of corpus linguistics developed, and surveys the major approaches to the use of corpus data. It uses a broad range of examples to show how corpus data has led to methodological and theoretical innovation in linguistics in general. Clear and detailed explanations lay out the key issues of method and theory in contemporary corpus linguistics. A structured and coherent narrative links the historical development of the field to current topics in 'mainstream' linguistics. Practical activities and questions for discussion at the end of each chapter encourage students to test their understanding of what they have read and an extensive glossary provides easy access to definitions of all technical terms used in the text.

TONY McENERY is Professor of English Language and Linguistics at Lancaster University.

ANDREW HARDIE is Lecturer in Corpus Linguistics at Lancaster University.

CAMBRIDGE TEXTBOOKS IN LINGUISTICS

Corpus Linguistics: Method, Theory and Practice

In this series:

S. C. LEVINSON *Pragmatics*
G. BROWN and G. YULE *Discourse Analysis*
R. HUDDLESTON *Introduction to the Grammar of English*
R. LASS *Phonology*
B. COMRIE *Tense*
W. KLEIN *Second Language Acquisition*
A. J. WOODS, P. FLETCHER and A. HUGHES *Statistics in Language Studies*
D. A. CRUSE *Lexical Semantics*
A. RADFORD *Transformational Grammar*
M. GARMAN *Psycholinguistics*
G. G. CORBETT *Gender*
H. J. GIEGERICH *English Phonology*
R. CANN *Formal Semantics*
J. LAVER *Principles of Phonetics*
F. R. PALMER *Grammatical Roles and Relations*
M. A. JONES *Foundations of French Syntax*
A. RADFORD *Syntactic Theory and the Structure of English: A Minimalist Approach*
R. D. VAN VALIN, JR, and R. J. LAPOLLA *Syntax: Structure, Meaning and Function*
A. DURANTI *Linguistic Anthropology*
A. CRUTTENDEN *Intonation* Second edition
J. K. CHAMBERS and P. TRUDGILL *Dialectology* Second edition
C. LYONS *Definiteness*
R. KAGER *Optimality Theory*
J. A. HOLM *An Introduction to Pidgins and Creoles*
G. G. CORBETT *Number*
C. J. EWEN and H. VAN DER HULST *The Phonological Structure of Words*
F. R. PALMER *Mood and Modality* Second edition
B. J. BLAKE *Case* Second edition
E. GUSSMAN *Phonology: Analysis and Theory*
M. YIP *Tone*
W. CROFT *Typology and Universals* Second edition
F. COULMAS *Writing Systems: An Introduction to their Linguistic Analysis*
P. J. HOPPER and E. C. TRAUGOTT *Grammaticalization* Second edition
L. WHITE *Second Language Acquisition and Universal Grammar*
I. PLAG *Word-Formation in English*
W. CROFT and A. CRUSE *Cognitive Linguistics*
A. SIEWIERSKA *Person*
A. RADFORD *Minimalist Syntax: Exploring the Structure of English*
D. BÜRING *Binding Theory*
M. BUTT *Theories of Case*
N. HORNSTEIN, J. NUÑES and K. GROHMANN *Understanding Minimalism*
B. C. LUST *Child Language: Acquisition and Growth*
G. G. CORBETT *Agreement*
J. C. L. INGRAM *Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders*
J. CLACKSON *Indo-European Linguistics: An Introduction*
M. ARIEL *Pragmatics and Grammar*
R. CANN, R. KEMPSON and E. GREGOROMICHELAKI *Semantics: An Introduction to Meaning in Language*
Y. MATRAS *Language Contact*
D. BIBER and S. CONRAD *Register, Genre and Style*
L. JEFFRIES and D. MCINTYRE *Stylistics*
R. HUDSON *An Introduction to Word Grammar*
M. L. MURPHY *Lexical Meaning*
J. M. MEISEL *First and Second Language Acquisition*
T. MCENERY and A. HARDIE *Corpus Linguistics: Method, Theory and Practice*

Corpus Linguistics

Method, Theory and Practice

TONY McENERY AND
ANDREW HARDIE

Lancaster University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

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

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521547369

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloguing in Publication data
McEnery, Tony, 1964–
Corpus linguistics : method, theory and practice / Tony McEnery, Andrew Hardie.
p. cm. – (Cambridge textbooks in linguistics)
Includes index.
ISBN 978-0-521-83851-1 (hardback)
1. Corpora (Linguistics) I. Hardie, Andrew. II. Title.
P128.C68M38 2011
410.1′88 – dc23 2011026519

ISBN 978-0-521-83851-1 Hardback
ISBN 978-0-521-54736-9 Paperback

Contents

Figures

Tables

Acknowledgements

This book could not have been written without the aid of many colleagues whose generous assistance we gratefully acknowledge. Svenja Adolphs, Karin Aijmer, Mark Davies, Costas Gabrielatos, Geoffrey Leech, Neil Millar and Richard Zhonghua Xiao all read part or all of the manuscript and offered useful criticisms and suggestions (although, as should go without saying, responsibility for the final text rests solely with us). We would also like to thank Eivind Torgerson for a helpful discussion regarding the similarity of corpus linguistics and sociolinguistics, and Ghada Mohamed for long talks on the topics of cluster analysis and the notion of an exhaustive linguistic feature tree (both of which fed into the discussion in Chapter 5). Others, too numerous to mention, among our colleagues and students pointed us towards relevant literature or gave us valuable sneak previews of books and research papers in press which we would not otherwise have been able to discuss here. We are thankful to them all. We are also grateful for the professionalism of the editorial staff at Cambridge University Press.

Preface

The title of this book is *Corpus Linguistics: Method, Theory and Practice*. As that title may suggest, it is about how corpus linguistics has developed and is employed as a methodology; the major theoretical issues that corpus linguists contend with today; and the problems that researchers using corpora must grapple with in practice, both within linguistics and across disciplines.

This captures in a nutshell, we like to think, what makes this book different from the many excellent introductory textbooks on corpus linguistics that have appeared since the mid-1990s. Our purpose in writing this book is not to introduce the very basics of procedures in corpus linguistics – we do not outline any step-by-step instructions describing how to go about investigating a corpus, and though we do occasionally give some example corpus analyses, we do not go into the details of how to deal with concordances, collocations, keywords and other common outputs from corpus tools. Other books have covered this ground comprehensively (e.g. Biber *et al*. 1998; Hunston 2002; Adolphs 2006; McEnery *et al*. 2006; Hoffmann *et al*. 2008), and we do not see any need to duplicate these accounts.

Instead, our aim in this textbook is to introduce, explain and in some cases problematise the most fundamental conceptual issues underlying the use of corpora, as well as reviewing what we see as the major trends of research using corpora to date. In the earlier part of the book, we will discuss corpus linguistics as a discipline, exploring high-level issues of practice that are of concern to corpus linguists, such as: features of corpus construction such as the notions of balance and representativeness; the development and exploitation of corpus tools; issues of copyright law and of research ethics; the role and limits of corpus annotation; the role of quantitative analysis; and so on. We will also review the research priorities and main contributions of a number of different schools and centres of corpus linguistics. In doing so, we will explore both their methodological apparatus and, where appropriate, the contributions to theory that they have sought to make. However, in later chapters we will move beyond the boundaries of corpus linguistics per se to consider the role that corpus methods now have across a range of types of linguistic investigation – including studies of language variation, language change, functional-cognitive linguistic theory, and psycholinguistics – and the various issues raised by the use of corpora in such enterprises. Accordingly we will argue that, although 'corpus linguistics' has clearly long had a separate existence, there is a very great degree of convergence between

corpus linguistics and these other aspects of linguistics. Corpus techniques tend no longer to be the preserve of a clearly delimited field of specialists, but rather have become a critical resource across linguistics as a whole (and beyond). Thus, we might argue that the future of the field is in 'corpus methods in linguistics' rather than 'corpus linguistics' standing independently.

This view, it must be noted, is not universally held. In particular, some scholars of the neo-Firthian school of corpus linguistics disagree entirely with it, as we will explain in our discussion of that tradition in Chapter 6. When covering matters such as this, which do not attract full consensus, we had two choices as authors. The first was to attempt to maintain neutrality, and to write the text without letting our opinions and theoretical and methodological preferences colour the account. The second was to acknowledge our perspective, and associated biases, frankly and explicitly, and to provide an account in which we explain and justify our views as best we can. It is this second approach which we have adopted. Our discussion of the different traditions of corpus research, of the wider use of corpora in linguistics and beyond, and of the future directions we see as desirable, thus amount to what might be called a position statement for our 'version' of corpus linguistics (to borrow the memorable phrasing of Teubert 2005). We make no pretence to neutrality, and our discussions of traditions in corpus linguistics other than our own may indeed characterise a given school of thought in terms that the researchers within that tradition would not necessarily agree with. The most notable case of this is that our discussion of neo-Firthian corpus linguistics is very much informed by our stance as non-neo-Firthian researchers. We urge the reader who is interested in understanding the full picture with regard to these debates to refer to accounts in which the scholars we discuss have dealt with these issues from their own perspective, with their different sets of preconceptions, opinions and biases; where appropriate we list such accounts in the suggestions for further reading provided in each chapter. In the case of neo-Firthian theory, for instance, we would not hesitate to recommend Tognini-Bonelli (2001) or Teubert and Čermáková (2004) as clear and readable presentations of the 'other side' of the argument that we make here in Chapters 6 and 8.

Our recommendations for further reading cover both the primary and the secondary literature, as appropriate. We have also included some study questions at the end of each chapter. These are divided into two groups. For readers who are interested in thinking further about some of the issues and problems that we cover, we suggest a number of *questions for discussion*. Alongside that, we also provide some *practical activities*. These activities provide practice in some of the key procedures that are needed in actually 'doing' corpus linguistics. All the practical activities assume that you have access to a reasonably large corpus and some corpus analysis software, but do not require any specific corpus. So if you have a copy of a suitable corpus on your computer, you can work with that data using any one of several different corpus tools to complete the exercise (some of which are available for free on the Internet). If not, then a number

of large, standard corpora – in several languages – can be accessed and analysed via the World Wide Web, using online interfaces which either are openly accessible, or offer freely available sign-in accounts. There are too many of these to list here (but see our discussion in Chapter 2). For English, we recommend BNCweb (http://bncweb.lancs.ac.uk/bncwebSignup) or the Brigham Young University online interface (http://corpus.byu.edu). While different software tools have different capabilities, the activities in this book are based on the core functions that all corpus tools make available.

The centrality of the Internet to the practice of corpus linguistics today means that it was necessary to make reference to websites and web services at various points in the book. We have tried to keep the number of web addresses in the text to a minimum, in the light of how variable these can be over relatively short periods of time. It is inevitable, however, that some of these web addresses will become outdated over time. For this reason, we have established a companion website for this book, where updated links will be made available where necessary. The website will also contain other supplementary material, including suggested answers to the study questions in each chapter. Importantly, this site also contains a large number of additional notes that were simply too numerous to include in the book. Accordingly, we suggest that you check the website after you read each chapter. The companion website address is www.cambridge.org/mcenery-hardie.

1 What is corpus linguistics?

1.1 Introduction

What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language (although, as we will see, at least one major school of corpus linguists does not agree with the characterisation of corpus linguistics as a methodology). The procedures themselves are still developing, and remain an unclearly delineated set – though some of them, such as concordancing, are well established and are viewed as central to the approach. Given these procedures, we can take a corpus-based approach to many areas of linguistics. Yet precisely because of this, as this book will show, corpus linguistics has the potential to reorient our entire approach to the study of language. It may refine and redefine a range of theories of language. It may also enable us to use theories of language which were at best difficult to explore prior to the development of corpora of suitable size and machines of sufficient power to exploit them. Importantly, the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language – theories which draw their inspiration from attested language use and the findings drawn from it. In this book, these impacts of corpus linguistics will be introduced, explored and evaluated.

Before exploring the impact of corpora on linguistics in general, however, let us return to the observation that corpus linguistics focuses upon a group of methods for studying language. This is an important observation, but needs to be qualified. Corpus linguistics is not a monolithic, consensually agreed set of methods and procedures for the exploration of language. While some generalisations can be made that characterise much of what is called 'corpus linguistics', it is very important to realise that corpus linguistics is a heterogeneous field. Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data. But let us first deal with the generalisations. We could reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions. The set of texts or *corpus* dealt

with is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe. It is the large scale of the data used that explains the use of machine-readable text. Unless we use a computer to read, search and manipulate the data, working with extremely large datasets is not feasible because of the time it would take a human analyst, or team of analysts, to search through the text. It is certainly extremely difficult to search such a large corpus by hand in a way which guarantees no error. The next generalisation follows from this observation: corpora are invariably exploited using tools which allow users to search through them rapidly and reliably. Some of these tools, namely concordancers, allow users to look at words in context.[1] Most such tools also allow the production of frequency data of some description, for example a word frequency list, which lists all words appearing in a corpus and specifies for each word how many times it occurs in that corpus. Concordances and frequency data exemplify respectively the two forms of analysis, namely qualitative and quantitative, that are equally important to corpus linguistics.

The importance of our findings from a corpus, whether quantitative or qualitative, depends on another general factor which applies to all types of corpus linguistics: the corpus data we select to explore a research question must be well matched to that research question. To some extent this is self-evident – a corpus is best used to answer a research question which it is well composed to address. To give an extreme example, there would be little point in exploring the noun classification system of Swahili by looking in a corpus of English newspaper texts. More subtly, we cannot (or can only with some caution) make general claims about the nature of a given language based on a corpus containing only one type of text or a limited number of types of text. Finally, and more subtly still, we must be aware that texts within a corpus that we assume to be homogeneous may, in fact, exhibit differences. For example, a collection of samples from a newspaper, even the same newspaper on the same day, may exhibit entirely predictable differences from one another – the sports section, for example, will draw on different lexis than the international news section. Users of a corpus must be aware of its internal variations, and researchers sometimes use statistical techniques to examine the degree of variability within a given corpus before using it (see Gries 2006c for an example of how to explore such variability within a corpus). The degree of homogeneity of a corpus is then another factor in determining how well matched that corpus is to particular research questions.

We have been discussing the features of texts within a corpus. It should be noted that the term *text* here denotes a file of machine-readable data. Typically in corpus linguistics these are in fact textual in form, so that each file represents, for instance, a newspaper article or an orthographic transcription of some spoken language. However, the computer files within a corpus do not need to be textual, and there are certainly examples nowadays of files of video data being used as corpus texts, as we will discuss in the next section.

This last point highlights a problem even with the very gross generalisations we have made so far – they are generally accurate, but we can very often find specific

examples that challenge them. For example, although we have said that corpus linguistics always uses machine-readable text, in fact, historically, much work was undertaken on corpora held in paper form; for example Fries (1952) produced a grammar of English based upon such a corpus. Also, while it is true that much research using corpus methods (e.g. McEnery 2005; Davies 2009b; Millar 2009; and many others) uses corpora of millions of words, there are others studies such as those of Ghadessy and Gao (2001) and McEnery and Kifle (2001) which, appropriately, use smaller, specialised corpora that might conceivably have been analysed by hand and eye. Nonetheless, despite the exceptions, the generalisations above characterise much of the work that can reasonably be described as corpus linguistics. Looking beyond these generalisations, research within the field can be divided on the basis of a number of criteria which discriminate quite sharply between types of work. The following features are those which, in our view, most typically distinguish different types of studies in corpus linguistics:

- Mode of communication;
- Corpus-based versus corpus-driven linguistics;
- Data collection regime;
- The use of annotated versus unannotated corpora;
- Total accountability versus data selection;
- Multilingual versus monolingual corpora.

Using these features, we can begin to work out a rough typology of corpus linguistic research, at least in terms of the principles underlying the use of corpora in such studies. Several of the later chapters of this book will be devoted to developing critical overviews of some of the types of corpus linguistics outlined within this typology, including the 'neo-Firthian' tradition (Chapter 6) and the variationist tradition (Chapter 5). However, in order to fully understand this typology, we clearly need to define the oppositions above in some detail.

1.2 Mode of communication

Corpora may encode language produced in any mode – for example, there are corpora of spoken language and there are corpora of written language. In addition, some video corpora record paralinguistic features such as gesture (Knight *et al*. 2009), and corpora of sign language have been constructed (Johnston and Schembri 2006; Crasborn 2008).

Corpora representing the written form of a language usually present the smallest technical challenge to construct. Until recently, encoding – and reliably representing on screen – writing systems other than the Roman alphabet was prone to error (Baker *et al*. 2000).[2] However, with the advent of Unicode (Unicode Consortium 2006), this problem is being consigned to history; Unicode allows computers to reliably store, exchange and display textual material in nearly all of

the writing systems of the world, both current and extinct. Written corpora can still be time-consuming and error-prone to produce in cases where the materials have to be either scanned or typed from printed original documents (this is particularly true for handwritten material – see Smith *et al*. 1998). However, as we will discuss later in this chapter, the increasing availability of a wide range of genres in machine-readable format for most major languages means that the construction of written corpora, except in the context of historical linguistic research, has never been easier.

Material for a spoken corpus, however, is time-consuming to gather and transcribe. Some material may be gathered from sources like the World Wide Web – for example, transcripts of parliamentary debates, called Hansard reports, are produced in the UK. These are readily accessible on the web.[3] Also, Hoffmann (2007a) has gathered transcripts of news broadcasts from the web to represent speech. However, transcripts such as these have not been designed as reliable materials for linguistic exploration of spoken language. Consequently there are 'serious hazards involved if transcripts that were made by non-linguists for purposes of their own are to be used for linguistic analysis' (Mollin 2007: 188). Mollin (2007: 208) outlines the dangers of using data such as Hansard, whose transcripts are known to make certain changes to what was actually said:

> Some of the changes are due to the fact that Hansard transforms conversation based on the here and now of the situation into a decontextualised report that is also understandable to the distant reader. Adding information on speakers and persons referred to . . . In addition, Hansard omits certain interpersonal and situational references, resulting in a reduction of the very typical parliamentary formulae, e.g. those of turn taking. The picture conveyed to the reader is one where MPs speak orderly one after the other without any apparent meta-comments on how and when to speak.

Given problems such as these, it is hardly surprising that spoken corpus data is more often produced by recording interactions and then transcribing them. Orthographic and/or phonemic transcriptions of spoken materials can be compiled into a corpus of speech which is searchable by computer. These transcriptions may be linked back systematically to the original recording through a process called time-alignment so that, through the computer, it is possible both to easily search a spoken corpus and to hear the portion of the recording that matches a particular search result. This is possible, for example, with the COLT corpus of London teenage speech (Stenström *et al*. 2002), the International Corpus of English British component (ICE-GB)[4] and the Origins of New Zealand English (ONZE) corpus (Fromont and Hay 2008). The orthographic form of a spoken corpus often normalises the form of the words in the text to standard spellings, meaning that orthographically transcribed material is rarely a reliable source of evidence for research into variation in pronunciation. Phonemically transcribed material is of much more use in this respect, though it tends to be most useful when variant forms can be searched for by reference to a standardised form, typically

the orthographic transcription. In this way, the differing phonemic transcriptions corresponding to a single standardised form in different contexts can be compared and contrasted (as is possible, for example, in the Spoken English Corpus; Knowles *et al*. 1996). An interesting issue arises when compiling or analysing a spoken corpus of a language for which there is no written form, or where the written form is not easily rendered in machine-readable form. In this case it may be necessary to rely on phonemic transcription alone, or to decide upon an orthographic transcription scheme that allows for recovery of forms which are equivalent yet vary phonetically.

Corpora which include gesture, either as the primary channel for language (as in sign language corpora) or as a means of communication parallel to speech, are relatively new. The integration of video analysis with textual analysis is clearly crucial for the development and use of such corpora. This may be achieved by using software packages, such as Eudico (Brugman *et al*. 2002) and the Digital Replay System,[5] which allow text, sound and video to be brought into alignment with one another for the purposes of searching and analysing the data. Given the advances in technology that were required in order to handle such data, it is unsurprising that corpus linguistic studies focusing on the visual medium are only just beginning to be undertaken on a truly large scale, for example investigating the relationship between gesture and speech (Carter and Adolphs 2008), or constructing large corpora of sign language material (Johnston and Schembri 2006). Novel analytical schemes have also needed to be developed to begin the process of analysing the video streams that form the raw data of these types of corpora (for examples see Wittenburg *et al*. 2002 and Knight *et al*. 2009).

These basic distinctions in mode of communication do not map simply onto corpus data – many corpora contain data from more than one mode, such as the British National Corpus (BNC; Aston and Burnard 1998), which contains both speech and writing. However, the medium of communication itself does produce a distinction which is linguistically meaningful. Large-scale work contrasting spoken and written language has led to a much deeper appreciation of how remarkably different the two can be, as is shown at the level of grammar, for example, by Biber *et al*. (1999) and Carter and McCarthy (1995). The differences are such that some linguists, notably Brazil (1995), have made the claim that the grammar of speech and that of writing are not merely distinct but entirely different (see section 4.6). Consequently, thinking about corpora in terms of mode of production is not just a matter of different data collection and technical issues; we would argue that it is, rather, linguistically a very real distinction.

1.3 Corpus-based versus corpus-driven linguistics

The difference between corpus-based and corpus-driven language study (to use the terms originally introduced by Tognini-Bonelli 2001) is a topic

that runs through this book. Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach to the use of corpus data in linguistics. Corpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that the corpus *itself* should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies its own theory of language (Tognini-Bonelli 2001: 84–5). This notion of corpus-driven linguistics is closely associated with the work of scholars we will refer to as 'neo-Firthians', which will be explored in depth in Chapter 6. In that chapter, we will also revisit and problematise the corpus-based versus corpus-driven distinction. For those who accept it, the corpus-based versus corpus-driven dichotomy creates a basic, binary distinction, under which most works of corpus linguistic research can be sorted into one or the other group. However, our own perspective rejects the notion that the corpus itself has a theoretical status, and thus also rejects the binary distinction between corpus-based and corpus-driven linguistics. From this point of view, *all* corpus linguistics can justly be described as corpus-based. This point of controversy will be explored in Chapter 6.

1.4 Data collection regimes

An important question follows from the observation that corpus studies should match their data to their research question. How can we ensure that the match is good enough? If we want to explore grammatical features in Modern English, we clearly need to match the data we use against the claims we wish to make. To make general claims about spoken English, we would require a suitable spoken dataset. The speech of one person alone is unlikely to provide a suitable basis for such generalisations. So corpus construction, and in particular data collection, emerges as a critical issue for corpus linguistics. Two broad approaches to the issue of choosing what data to collect have emerged: the *monitor corpus* approach (see Sinclair 1991: 24–6), where the corpus continually expands to include more and more texts over time; and the *balanced corpus* or *sample corpus* approach (see Biber 1993 and Leech 2007), where a careful sample corpus, reflecting the language as it exists at a given point in time, is constructed according to a specific sampling frame.

1.4.1 Monitor corpora

The monitor corpus approach, proposed most notably by John Sinclair, seeks to develop a dataset which grows in size over time and which contains a variety of materials. The relative proportions of the different types of materials may vary over time. Monitor corpora could be said to balance any need to be

precise about the composition of a corpus against sheer size – as the corpus grows, we might assume that any skew in the data naturally self-corrects, since there is no *consistent* skew in the data input. The Bank of English (BoE), developed at the University of Birmingham, is the best-known example of a monitor corpus. The BoE was started in the 1980s (Hunston 2002: 15) and has been continually expanded since that time. At the time of writing, the corpus contains over half a billion words, organised into a general English section (450 million words) and a section containing corpus materials of use in language pedagogy (56 million words). The BoE represents one approach to the monitor corpus; the Corpus of Contemporary American English (COCA; Davies 2009b) represents another.[6] COCA expands over time like a monitor corpus, yet it does so according to a much more explicit design than the BoE. Each extra section added to COCA complies to the same, breakdown of text-varieties. Arguably, this corpus represents something of a halfway house between the sample corpus approach and the monitor corpus approach – a monitor corpus that proceeds according to a sampling frame and regular sampling regime.

While the BoE and COCA are impressive in scale, there is arguably a much larger monitor corpus under construction that covers a wide range of languages and contains a growing record of those languages over time – the World Wide Web.

1.4.2 The Web as Corpus

The concept of *Web as Corpus* (Kilgarriff and Grefenstette 2003) is very similar in many ways to the idea of the monitor corpus. It takes as its starting point a massive collection of data that is ever-growing, and uses it for the study of language (see, for example, the web-based study of antonyms by Jones *et al*. 2007, as a good example of the use of the web as a corpus). As well as using standard search engines such as Google to explore the web as a corpus, researchers have also developed interfaces specifically designed to support this use of the web, such as WebCorp (Renouf 2003). The Web as Corpus approach has some specific problems. In contrast to most corpora, the web is a mixture of carefully prepared and edited texts, and what might charitably be termed 'casually prepared' material. The content of the web is also not divided by genre – hence the material returned from a web search tends to be an undifferentiated mass, which may require a great deal of processing to sort into meaningful groups of texts. In addition, there is little doubt that the many texts on the web contain errors of all sorts. For example, while writing this book we typed *receive* and *recieve* into Google – *receive* scored 300,000,000 hits, *recieve* scored 8,670,000 hits. This of course may prove useful – if you wish to investigate common spelling errors, for example. Data like this might also be the basis for a very interesting study in support of spelling reform. However, if this isn't the sort of thing you are interested in, such errors in the data may well provide unwelcome noise when the analyst approaches the web as a corpus. Given that this kind of noise exists at all

levels of language on the web, it represents a significant issue that the users of the web as a corpus must address. Nonetheless, the web does undoubtedly provide a substantial volume of data which can be selected and prepared to produce corpora suitable for a wide variety of purposes.

By way of illustration, if you wanted to examine the rather loaded phrase *swanning around* in the BNC, you would find only 13 examples of it on which to base your observations. Using Google, we recovered 32,300 examples of texts containing this phrase. Admittedly, those thousands of examples would need to be sorted and sifted before they could be used to explore the phrase reliably. However, there is little doubt that the thousands of examples from Google would allow a more nuanced investigation of this particular phrase than the dozen or so examples in the BNC. So the web is a useful and readily available source of evidence, which can be invaluable in cases where you need a large quantity of data in order to deal with a low frequency of occurrence. However, there is a problem associated with this opportunity: for frequent words or phrases, the number of examples returned by a web search engine may simply be overwhelming, and a good deal of data may have to be discarded. This should be done in accordance with some heuristic which ideally should be applied consistently across all analyses. For example, if we study frequent words like *receive* using the web as a corpus, we may want to study only the first 100 examples that are returned. If we do this for one word in the study, then we should do so for *all* words in that study, where necessary. Another problem exists with all studies based on web data that is not downloaded and archived appropriately: the web is forever changing. It is difficult to replicate a study done on the web four years ago, for example, as the web will have changed significantly. Given the importance of replicability in experimental procedures (which we will discuss in section 1.6.1), this is an obvious and pressing drawback to the Web as Corpus approach.

1.4.3 The sample corpus approach

In contrast to monitor corpora, balanced corpora, also known as *sample corpora*, try to represent a particular type of language over a specific span of time. In doing so, they seek to be *balanced* and *representative* within a particular *sampling frame* which defines the type of language, the *population*, that we would like to characterise. The population is the notional space within which language is being sampled. So, for example, if we want to look at the language of service interactions in shops in the UK in the late 1990s, the sampling frame is clear – we would only accept data into our corpus which represents service interactions in UK shops in the 1990s. However, if we only collected data gathered in coffee shops, we would not get a balanced set of data for that population – relatively context-specific lexis, such as *latte* and *frapuccino*, would be likely to occur much more frequently than they do in service interactions in general. Phrases which are typical of other kinds of service interactions, such as *Should I wrap that for you?*, might not occur at all. Following the principle of *balance*, we would seek

to characterise the *range* of shops whose language we wanted to sample, and collect data evenly from across that range.

Even if we decided we were only interested in bookshops, coffee shops and supermarkets, we might still wish to ensure that the shops sampled from were in some sense *typical*, and that we gathered data from them in such a way as to avoid introducing skew into our dataset. So, we might care to ensure that we did not sample from bookshops which sell only antiquarian books, if we were concerned that the interactions there could be atypical of bookshops in general. Similarly, we might want to ensure that the proportions of data in our corpus reflect, in some way, the numbers of each type of interaction of interest that actually occur. If we had 90 per cent of our data from bookshops, 8 per cent from coffee shops and 2 per cent from supermarkets, when we know that there are a hundred supermarkets for every bookshop, we might well feel that our corpus design was less than ideal. We would have to choose the locations to sample from, and the relative proportions of different types of data to collect, with the aim of achieving *representativeness* for the data in a corpus. Of course, this simple example presents only one approach to representativeness; see Leech (2007) for a critical exploration of this concept.

Corpora which seek balance and representativeness within a given sampling frame are *snapshot* corpora. A good example of a snapshot corpus is the Lancaster-Oslo/Bergen (LOB) corpus. This represents a 'snapshot' of the standard written form of modern British English in the early 1960s. Table 1.1 gives the sampling frame within which the data for the LOB corpus was gathered.

For each category, samples of data were gathered, with each sample being of roughly similar length (2,000 words). The samples were taken from a variety of sources within each broad sampling domain. The resulting corpus is 1 million words in size. The LOB corpus demonstrates how a snapshot corpus, used in concert with corpora constructed using the same sampling frame, can allow us to undertake a wide range of contrasts and comparisons. The same sampling frame used for LOB has also been used to collect corpora of written British English at spaced intervals (mostly of thirty years) through the twentieth and early twenty-first centuries. This allows the effects of diachronic change to be studied in this variety of English (see Leech 2004; Baker 2009; Leech *et al.* 2009). This approach to exploring diachronic change is analogous to stop-motion photography – slow-moving changes become visible when a snapshot is taken at discontinuous intervals. It is also possible to study diachronic change with a large monitor corpus, though different techniques may be needed in order to capture slow-moving change over time. Using snapshot corpora, we can also look at synchronic differences in varieties of English. The LOB sampling frame was adopted from one originally developed to construct a corpus of written American English from 1961, the Brown Corpus (Francis and Kučera 1964; Kučera and Francis 1967) – so comparing LOB and Brown, we can investigate differences in the two language varieties while controlling for sampling and the effects of diachronic change. We will return to the study of synchronic and diachronic variation using the LOB and Brown corpora in section 5.3.

Table 1.1 *The LOB Corpus Sampling Frame (after Hofland and Johansson 1982: 2)*

Category mnemonic	Description	Number of text samples in this category
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	38
F	Popular lore	44
G	Belles lettres, biography, essays	77
H	Miscellaneous (government documents, foundation reports, industry reports, college, catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
Total		500

1.4.4 Balance, representativeness and comparability

Balance, representativeness and comparability are ideals which corpus builders strive for but rarely, if ever, attain. In truth, the measures of balance and representativeness are matters of degree. Váradi (2001) has been critical of the failure of corpus linguists to fully define and realise a balanced and representative corpus. Even proposals, such as those of Biber (1993), to produce empirically determined representative corpora have not actually been pursued. Biber's proposal for representativeness to be realised by measuring internal variation within a corpus – i.e. a corpus is representative if it fully captures the variability of a language – has yet to be adopted in practice. It is also only one of many potential definitions of representativeness, as Leech (2007) points out. However, though balance and representativeness remain largely heuristic notions, decided on the basis of the judgement of linguists when they are building a corpus, this does not mean to say that the concepts are of no value. Similarly, while some corpora designed to be comparable to each other can clearly make a claim for balance and representativeness, others may only do so to a degree. Leech (2007: 141–3) usefully summarises a series of problems encountered in building comparable corpora of British English to explore diachronic variation: notably,

problems relating to the evolution over time of the genres that are balanced in those corpora. The changing nature of genre makes claims of comparability when looking at diachronic variation much more tendentious than similar claims for the synchronic Brown/LOB comparison, for example. As Leech (2007: 143–4) notes, the debate around balance, representativeness and comparability might lead researchers:

> to reject these concepts as being ill-defined, problematic and unattainable. My attitude is different from this . . . these are important considerations, and even if we cannot achieve them 100 per cent, we should not abandon the attempt to define and achieve them. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design. It is best to recognise that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than abandon them altogether.

There is little doubt that, as the corpus approach to language develops, the concepts of balance and representativeness will undergo further critical scrutiny. This in turn should lead to incrementally better definitions of these terms.

1.4.5 'Opportunistic' corpora and minority and endangered languages

The monitor versus snapshot corpus distinction provides us with a ready framework for categorising corpora which make some claim to represent a particular language in general. However, it also must be noted that there are many collections of data, reasonably described as corpora, which do not necessarily match the description of either a monitor or a snapshot corpus comfortably. Such corpora are best described as *opportunistic* corpora. These corpora make no pretension to adhere to a rigorous sampling frame, nor do they aspire to deal with issues of skew by the collection of an ever-larger body of data, as monitor corpora may. Rather, they represent nothing more nor less than the data that it was possible to gather for a specific task. Sometimes technical restrictions prevent the collection of large volumes of data to populate some idealised sampling frame. This was particularly true prior to the widespread introduction of electronic publishing and the general availability of electronic text in a range of languages on the web. Some early corpora were not built along principled lines according to the demands of a specific research question; rather, they were constructed using whatever relevant material could be accessed in electronic form. Corpora such as the American Printing House for the Blind corpus (Black *et al*. 1993) and the Hansard Corpus (Berger *et al*. 1994) were built in order to exploit materials from what were, at the time, two of the very few text producers who created machine-readable versions of texts. This problem clearly no longer generally applies to English or most other major languages, but it still persists for some languages.

It is likely that for languages with a written form, more and more machine-readable textual material will become available over time, allowing them to be readily studied. Consider the general division of languages into four broad types suggested by McEnery and Ostler (2000):

1. Official majority languages (e.g. English in the UK, Portuguese in Portugal).
2. Official minority languages (e.g. Welsh in the UK).
3. Unofficial languages (both large, e.g. Kurdish in Turkey, and relatively small, e.g. Sylheti in the UK).
4. Endangered languages (e.g. Guugu Yimidhirr in Australia).

It is fair to say that types 1 and 2 are better supplied with corpus data than 3 and 4 for a range of non-linguistic reasons. Official languages typically have governments with money associated with them. These governments typically publish material in the official language, often on the web. They also, at times, fund corpus-building projects. Unofficial languages suffer from a lack of official recognition, and hence state funding. Furthermore, if the language is associated with an oppressed group, the language itself may be suppressed. The issue with endangered languages is obvious – very few speakers producing little material relative to the larger languages. It may also be the case that endangered languages are also suppressed, making their situation yet worse.

A significant problem arises in the context of analysts approaching spoken data in particular: converting spoken recordings into machine-readable transcriptions is a very time-consuming task. This in itself means that, without significant financial support or plenty of available time, some analysts choose to work on small datasets when much larger datasets would arguably be more appropriate for their task. Analysts may feel, rightly given the resources available, that working with a small sample may be sufficient for their purposes, and that while a larger dataset might yield slightly different results, they face the prospect of 'a huge amount of work and planning for very small returns' (Holmes 1996: 168). A researcher must at times be guided by pragmatism.

Finally, even with a huge amount of work and planning, it may simply be impossible to build an ideal corpus for a given language – if the language is dead or dying and the material to construct a large, balanced corpus is not available and simply never will be. To consider an extreme example, the Indus Valley civilisations based around Harappa and Mohenjo-daro flourished between approximately 2,500 and 1,900 BCE. The total stock of written material that remains to represent the language used by that civilisation consists of 3,700 inscribed objects (Robinson 2009: 268). It is unlikely that future archaeological digs will significantly alter the extent of this stock of text. If we want to build a corpus from these objects, perhaps to try to decode this as yet undeciphered script, the amount of material to draw on is quite finite – the language is dead and the writing system is no longer used. In all likelihood we have the great majority of surviving 'texts' in our possession already. No native speaker of the

language of the Indus Valley will ever again exist to produce more texts using this writing system. When dealing with an extinct language for which a greater body of literature survives, such as Classical Latin, Gothic or Old English, our situation is different in degree but not in kind from the Indus Valley case: our only choice in building a corpus is to select some or all of the texts that have made it through the centuries.

In summary, while the notions of monitor and snapshot corpora provide us with relatively idealised models of corpus construction, it should be noted, and accepted, that the corpora that we use and construct must sometimes be determined by pragmatic considerations.

1.5 Annotated versus unannotated corpora

A further way in which studies in corpus linguistics vary relates to whether or not linguistic analyses are encoded in the corpus data itself. Such encoding, called *corpus annotation*, may be achieved either by editing the data to include within it some analysis, or by having the analysis stored separately but linked in to the data. For example, we may wish to annotate a corpus to show parts of speech, assigning to each word the grammatical category we claim it has in its context. So, for example, when we see the word *talk* in the sentence *I heard John's talk and it was the same old thing*, we would assign it the category 'noun' in that context. In doing so, we might edit the text directly, assigning some mnemonic code (such as N) to make it clear that in this case the word is a noun. In a simple case, we may just attach the mnemonic code to the word in question with an underscore – *talk_N*.[7] Rather than edit the text directly, however, it is also possible to store annotations like this separately from the data itself, using computer programs to combine, integrate and disentangle the text and annotations as the analyst desires. This so-called 'stand-off' annotation is preferred by some analysts (e.g. Thompson and McKelvie 1997). However, given a systematic encoding of annotations directly in a corpus, it is a trivial matter to remove them if desired, so the arguments in favour of stand-off annotation seem to boil down more to a question of methodical neatness or elegance rather than denoting anything fundamental in nature.[8]

While the phrase *corpus annotation* may be unfamiliar to some linguists, the basic operation it describes is not – it is directly analogous to the analyses of data that have been done using hand, eye and pen for decades. Corpus annotation is, then, a commonplace of linguistics. If it varies from usual practice at all, it is in the scale on which it is applied. In Chomsky (1965), twenty-four invented sentences are analysed; in the parsed version of LOB, a million words are annotated with parse trees. Nonetheless, it is important to note that, setting scale aside, corpus annotation is largely the process of providing – in a systematic and accessible form – those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they worked with.

On the basis of this somewhat brief description of corpus annotation, a reader would be forgiven for thinking that the distinction between annotated and unannotated corpora is based simply on whether or not the corpus has been analysed in a particular way *yet*. Those corpora which have already been analysed in some way are annotated, those which have yet to be analysed are not. This distinction in itself, however, is so trivial that it would hardly constitute a major dimension along which research in corpus linguistics can vary. What makes this dimension salient is the fact that some linguists object to annotation – either per se, or when undertaken manually rather than automatically by a computer. Opposition to annotation is typically associated with neo-Firthian corpus linguistics and the corpus-driven approach, as will be discussed in Chapter 6. However, in brief, arguments against annotation are largely predicated upon the purity of the corpus texts themselves, with the analyses being viewed as a form of impurity. This is because they impose an analysis on the users of the data, but also because the annotations themselves may be inaccurate or inconsistent (Sinclair 1992). Such claims are interesting because, as has been noted, corpus annotation is the manifestation within the sphere of corpus linguistics of processes of analysis that are common in most areas of linguistics. To identify problems with accuracy and consistency in corpus annotation is, in principle at least, to identify flaws with analytical procedures across the whole of linguistics. It is because of the issues of accuracy and consistency, in particular, that some linguists prefer to use unannotated corpora. But this does not mean to say that such linguists do not analyse the data they use; rather, it means that they leave no systematic record of either their analysis or their errors which can easily and readily be tied back to the corpus data itself.

1.6 Total accountability versus data selection

So far we have focused on ways in which corpora vary in their design. Corpora may also vary, however, in how they are used by the analysts who exploit them. A key difference here is the contrast between *total accountability* and *data selection*.

1.6.1 Total accountability, falsifiability and replicability

It has been argued that a significant advantage of using corpora is that corpora allow analysts to approach the study of language within the context of the scientific method (Leech 1992). A core principle of Leech's approach within this framework is total accountability (Leech 1992: 112). If you approach a corpus with a specific theory in mind, it can be easy to unintentionally focus on and pull out only the examples from the corpus that support the theory (this is technically called a *confirmation bias*). But the theory can never be shown to be false by such

an approach, even in principle. As such, this approach runs counter to one of the key features of the scientific method identified by Popper ([1934] 2006: 18), namely *falsifiability*. The principle of total accountability is, simply, that we *must not* select a favourable subset of the data in this way. When approaching the corpus with a hypothesis, one way of satisfying falsifiability is to use the entire corpus – and all relevant evidence emerging from analysis of the corpus – to test the hypothesis. This principle is the reason for the quantitative nature of many corpus-based methods. Minimally, however, where there is too much evidence for using the entire corpus to be practical, the analyst must at least, as Leech suggests, avoid conscious selection of data. Short of using the corpus in its totality, total accountability can in principle be preserved by using an unbiased (e.g. randomised) subsample of the examples in the corpus. If it were permissible, in corpus research, to filter out or ignore examples or statistics from the corpus that do not fit the hypothesis under investigation, then the corpus could support such a bewildering variety of potentially contradictory hypotheses that the use of corpus data would be fatally undermined. To put it simply, there should be no motivated selection of examples to favour those examples that fit the hypothesis, and no screening out of inconvenient examples. Such a statement represents an ideal for the use of corpus data that most would find difficult to challenge.

However, there is a criticism to be levelled at such an approach: the corpus *itself* is necessarily a finite subset of a much larger (and in principle non-finite) entity, language. So the corpus itself represents a selection and screening of data. Therefore, any claim of total accountability in corpus linguistics must be moderated. We can only seek total accountability relative to the dataset that we are using, not to the entirety of language itself. This criticism is not, of course, unique to linguistics. An obvious parallel is astronomy, where astronomers theorise on the basis of the subset of the Universe that is visible to them. They expand their dataset over time, and each generation of astronomers seeks to falsify the findings of the previous generations of astronomers as they push forward the boundaries of the field. A very similar model is developing in linguistics, now that it has become possible to expose linguistic theories to testing by large-scale observation. Based on this analogy, we can say that, like an astronomer, a corpus linguist can work in accordance with the scientific method, and produce potentially falsifiable results, while not being totally accountable in the strictest sense.

But moderating the claim of total accountability in the light of the finite size of the corpus does raise one troubling possibility. An analyst may, by chance or design, construct a dataset that misrepresents the language such that the analysis of this dataset supports a faulty theory. While we must be mindful of this possibility, an analogy with astronomy may help once again. Let us imagine an astronomer, at some point in the past, seeking to develop a model of moons based on data from the Earth, Mars and Jupiter. They then conclude, from that dataset, that all planets have moons. The problem here is with the dataset – it has unwittingly been drawn from a set of planets which happen to have moons. If Mercury or Venus, which lack moons, had been in the dataset, the conclusion would have been

different. The answer to the problem in astronomy is the same as in linguistics, and emerges from another key feature of the scientific method: *replicability*. A result is considered replicable if a reapplication of the methods that led to it consistently produces the same result. This process of checking and rechecking may be done with the same dataset or it may be done with new datasets. In Popper's theory, falsifiability is of higher priority than replicability as a key to verification in the scientific method. The ability to replicate a result, whether experimental or observational, is, nonetheless, still clearly central to scientific practice. In all the sciences, new results are typically considered provisional until they are known to be replicable – and in many cases, it is precisely through that process of continuous checking of results as theories develop and expand that falsifiability is achieved.

Like the natural sciences, corpus linguistics has in many cases appealed to the notion of the replicable result for credibility (see Doyle 2005 for a good critical overview of the engagement of corpus linguistics with replicability). In particular, replicability helps us address the problem of the limited dataset outlined above. Attempts to replicate the astronomical result that all planets have moons will, eventually, find that in a wider dataset of planets, the rule does not hold. Similarly, an incorrect or incomplete result that stems from the finite size of a corpus is likely to be found out when corpus linguists recheck that result against other datasets. So as long as this process of checking and replication runs its course, and given sufficient time and data, bias in the data of the sort we have outlined is routinely discovered and removed. There is evidence of this happening already in linguistics in general, and corpus linguistics in particular. A good example of work undertaken on one corpus being revised when further corpus data became available is Leech's (1971, 2004a) work on non-finite verbs (see also section 2.2). In sum, then, total accountability to the data at hand ensures that our claims meet the standard of falsifiability; total accountability to *other* data in the process of checking and rechecking ensures that they meet the standard of replicability; and the combination of falsifiability and replication can make us increasingly confident in the validity of corpus linguistics as an empirical, scientific enterprise.

1.6.2 Data selection – not (necessarily) a bad thing

Given what has been said about total accountability, you may wonder that analysts would ever approach a corpus seeking a single example, or a subset of carefully selected examples. Not only do some analysts do just that, in certain circumstances it may actually be the right thing to do. Indeed, in an important sense, approaching a corpus in search of a specific type of result may be entirely in line with the scientific method. Most importantly, we may seek in a corpus a specific example which, in itself, falsifies a hypothesis – thereby making the totality of the data in some sense irrelevant. One example alone may be enough to falsify a claim. In a corpus of a million sentences, the one sentence that

does not conform to a hypothesis is the only sentence that really matters for considering the hypothesis in question. This may be illustrated by returning to our astronomy parallel. Given the hypothesis that all planets have moons, if we have data available from a thousand planets, the fact that 999 of them have moons is not as important – from the point of view of defending the hypothesis – as the fact that one planet has no moons at all. Likewise, if the hypothesis we are looking at is that some particular linguistic form never occurs, then the only part of the corpus that is really relevant is the part where that linguistic form *does* occur, thus falsifying the hypothesis. To put this in general terms, a single example may falsify a hypothesis, leading to the revision, or abandonment, of that specific hypothesis. In that sense, approaching a corpus to find a single example is entirely consistent with both the scientific method and with the principle of total accountability.

A more contentious manifestation of utilising only selected parts of a corpus arises when researchers use the corpus simply as a bank of examples to illustrate a theory they are developing – this is sometimes called *corpus-informed* research. This clearly does run counter to the scientific method, insofar as there is no attempt to account for the rest of the (potentially falsifying) evidence in the corpus. However, some researchers have articulated an interesting motivation for using corpora in such a fashion. The premise is not unlike that which drives corpus linguistics to validate and revalidate hypotheses – namely, that the corpus is finite, but language is not. Some researchers argue that corpora, while a helpful guide or source of examples, cannot give sufficient access to language to the extent that so-called 'qualitative' approaches to the data should be abandoned. A good example of this has emerged in Critical Discourse Analysis (CDA).

CDA has traditionally been approached by the detailed analysis of single texts or small numbers of texts. On the basis of that detailed analysis, general claims about the use of language in society have then been made. Over time, as evidence from the analysis of individual texts has accumulated, overarching theories of how discourses work in society have emerged; and generic claims about the structure and nature of such discourse, focused, for example, on specific words or classes of word such as pronouns, have been made. These general observations, based on a small number of texts, have been exploited within an overarching theoretical framework based upon some theory of power relations. Since the mid-1990s, attempts have been made to integrate the general methodological approach of corpus linguistics with CDA by researchers such as Mautner (see Hardt-Mautner 1995, 2000; Mautner 2009), Koller and Mautner (2004), O'Halloran and Coffin (2004), Baker (2004, 2006, 2009) and Orpin (2005). A general issue with most of these attempts at integration has been one of balance – studies have tended either to focus mainly on either corpus linguistics or CDA at the expense of the other. Corpus-based studies may have explored discourse and its relation to power, but they have typically not been explicitly informed by CDA theory and its traditional methods, or else they have not aimed to contribute to a particular discourse-oriented theory (e.g. Stubbs 1994; Krishnamurthy 1996). Similarly,

CDA researchers have at times used data and techniques which are undoubtedly inspired by work in corpus linguistics, but have not sought to engage fully with the corpus approach (e.g. Fairclough 2000; Kovàcs and Wodak 2003). Research which is principally CDA-oriented tends to make limited or casual use of a corpus or corpus-based techniques. Sometimes, the corpus is used simply as a repository of examples (e.g. Flowerdew 1997) and no effort is made to apply the principle of total accountability that is generally accepted within corpus linguistics. Also, CDA studies making use of corpora have in general tended to avoid carrying out quantitative analyses beyond the simplest of descriptive statistics (see also Stubbs 1997: 104), preferring to undertake qualitative analyses using concordances.

Why do some researchers in CDA only engage minimally with corpus data? An important argument presented by such researchers relates to the depth of analysis that they want using the data they have – they wish to undertake a detailed analysis of a small amount of data, taking into account not just the text itself, but also the social context in which it was produced and the social context in which it was interpreted. This work is so labour-intensive that a large-scale study using the corpus may not be possible.[9] This argument has some weight. However, there is also the possibility of striking a balance where the corpus data itself is used in the framework of total accountability, but the detailed analysis is reserved for a subset of the data, once those hypotheses that are testable in practical terms on the whole corpus have been tested (KhosraviNik 2009). Nonetheless, it is still the case that many researchers prefer to work with small amounts of data in detail rather than engage with large corpora.

1.7 Monolingual versus multilingual corpora

Another obvious way in which corpora vary relates to the number of languages represented in the corpus.[10] Many corpora are monolingual in the sense that, while they may represent a range of varieties and genres of a particular language, they are nonetheless limited to that one language. So the International Corpus of English (ICE; see also section 4.2), for example, is a large monolingual corpus – it represents one language, English, though it allows linguists to compare and contrast a number of international varieties of that language. Monolingualism in corpora may be a matter of degree rather than an absolute. The BNC, for example, does contain some foreign words and speech produced by non-native English speakers (Aston and Burnard 1998: 127). However, the appearance of such data in the BNC does not reflect its primary purpose, which is to represent modern British English. The fact that some material in a language other than English was inadvertently collected does not mean that we should regard this corpus as anything other than what it claims to be – a monolingual corpus of English. However, the BNC could conceivably be considered (part of) a multilingual corpus if it were brought together with

a range of other corpora, of comparable size, scale and sampling frame, which happen to represent languages other than English. In order to understand this point, we need to consider the variety of multilingual corpora available.

When we refer to a corpus involving more than one language as a multilingual corpus, we are using the term *multilingual* in a broad sense to indicate 'two or more languages'; in a narrower sense, a multilingual corpus must involve at least three languages, while those involving only two languages are conventionally referred to as *bilingual* corpora. Given that corpora involving more than one language are a relatively new phenomenon, with most research hailing from the early 1990s (e.g. the English-Norwegian Parallel Corpus or ENPC; see Johansson and Hofland 1994), it is unsurprising to discover that there is some confusion surrounding the terminology used in relation to these corpora. Generally, there are three types of corpora involving more than one language:

- Type A: Source texts in one language plus translations into one or more other languages, e.g. the Canadian Hansard (Brown *et al*. 1991), CRATER (McEnery and Oakes 1995; McEnery *et al*. 1997).
- Type B: Pairs or groups of monolingual corpora designed using the same sampling frame, e.g. the Aarhus corpus of contract law (Faber and Lauridsen 1991), the Lancaster Corpus of Mandarin Chinese (McEnery *et al*. 2003), which uses the same sampling frame as LOB and Brown.
- Type C: A combination of A and B, e.g. the ENPC (Johansson and Hofland 1994), the EMILLE corpora (Baker *et al*. 2004).[11]

Different terms have been used to describe these types of corpora. For Aijmer *et al*. (1996) and Granger (1996: 38), type A is a *translation corpus* whereas type B is a *parallel corpus*; for Baker (1993: 248; 1995, 1999), McEnery and Wilson (2001: 70) and Hunston (2002: 15), type A is a *parallel corpus* whereas type B is a *comparable corpus*; and for Johansson and Hofland (1994) and Johansson (1998: 4–5), the term *parallel corpus* applies to both types – A and B. Barlow (1995, 2000: 110) certainly interpreted a 'parallel' corpus as type A when he developed the *ParaConc* corpus tool. It is clear that some confusion centres around the term *parallel*.

When we define different types of multilingual corpora, we can use different criteria, for example the number of languages involved and the content or the form of the corpus. But when a criterion is decided upon, the same criterion must be used consistently. For example, we can say a corpus is monolingual, bilingual or multilingual if we take the number of languages involved as the criterion for definition. We can also say a corpus is a translation (L2) or a non-translation (L1) corpus – type A or type B in the framework above – if the criterion of corpus content is used. But if we choose to define corpus types by the criterion of corpus form, we must use it consistently. Then we can say a corpus is parallel if the corpus contains source texts and translations in parallel, or it is a comparable corpus if its subcorpora are comparable by applying the same sampling frame. It

is illogical, however, to refer to corpora of type A as 'translation' corpora by the criterion of content while referring to corpora of type B as 'comparable' corpora by the criterion of form. Consequently, in this book, we will follow Baker's terminology in referring to type A as parallel corpora and type B as comparable corpora. As type C is a mixture of the two, corpora of this type should be referred to as comparable corpora in a strict sense.

A comparable corpus can thus be defined as a corpus containing components that are collected using the same sampling method, e.g. the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*. We previously observed that the BNC could conceivably become a sub-part of a comparable corpus if corpora similar to the BNC were collected in a range of languages. The resulting collection of corpora could be viewed as a multilingual corpus. However, the sub-parts of this multilingual corpus could also be considered monolingual corpora in their own right. Where there is an equivalence of sampling frames between corpora in different languages, they may be viewed and used as either monolingual or multilingual corpora as necessary. The subcorpora of a comparable corpus are not translations of each other. Rather, their comparability lies in the similarity of their sampling frames.

By contrast, a parallel corpus can most easily be defined as a corpus that contains native language (L1) source texts and their (L2) translations. This definition assumes that parallel corpora are unidirectional (e.g. from English into Chinese or from Chinese into English, but not both). This is currently the most common form of parallel corpus; for instance, the CRATER and EMILLE corpora already mentioned, as well as MULTEXT and P-ACTRES (Izquierdo *et al*. 2008), are unidirectional. However, there are some parallel bidirectional corpora, such as the Portuguese/English COMPARA corpus (Frankenberg-Garcia and Santos 2003),[12] the Nepali/English parallel section of the Nepali National Corpus (Yadava *et al*. 2008) and the English Swedish Parallel Corpus (Altenberg and Aijmer 2000); and there also exist multidirectional corpora (see, for example, the ECC-TEC corpus, Laviosa 2002). Arguably texts which are produced simultaneously in different languages (e.g. EU and UN regulations) can also be classed as parallel data (Hunston 2002: 15).

While parallel and comparable corpora are supposed to be used for different purposes (typically translation research and contrastive studies respectively; see Johansson 2007), the two are also designed with different focuses. For a comparable corpus, the sampling frame is essential. All the components must match with each other in terms of what types of texts they sample, in what proportions, from what periods. For the translated texts in a parallel corpus, the sampling frame is irrelevant, because all of the corpus components are exact translations of each other. Once the source texts have been selected in the first place, there is no need to worry about the sampling frame in the other language. However, this does not mean that the construction of parallel corpora is easier. For a parallel corpus to be useful, an essential step is to *align* the source texts and their

translations, annotating the correspondences between the two at the sentence or word level (see Oakes and McEnery 2000 for an overview). While this would ideally be accomplished using a computer program rather than manual analysis, the automatic alignment of parallel corpora is not a trivial task for some language pairs (Piao 2000, 2002).

1.8 Summary

By looking at a series of defining features in corpus linguistics, this chapter has explored ways in which the construction and use of corpora vary. In doing so, we have highlighted some of the differences that exist between linguists in their use – and basic conception – of corpus linguistics. In the two chapters that follow, we will shift the focus of our discussion to consider a range of more practical matters that corpus linguists face – how to annotate corpus data, how to analyse it and how to employ statistical techniques. We will also consider some of the constraints placed on corpus research by legal and ethical considerations. Throughout this discussion, however, key themes of this chapter will be returned to as they impact upon these practical issues. For example, the decision on whether to annotate or not is an important issue of principle as well as being a practical consideration. Likewise, the World Wide Web presents analysts with specific legal challenges, and the collection of spontaneous speech may bring with it significant ethical issues. So this chapter has raised some core issues which will resurface in a number of ways not only in the two chapters that follow, but also throughout the rest of this book.

Further reading

There is a growing body of books that deal in general with the topic of corpus linguistics. For those readers particularly interested in an approach to corpus linguistics which focuses upon genre analysis and textual variation, Biber *et al*. (1998) is both comprehensive and strongly recommended. With a somewhat different focus, Kennedy (1998) covers in some detail how corpus linguistics and language teaching in particular have intersected. For a general overview of corpus linguistics, with a discussion of its fall from favour in the mid-twentieth century, McEnery and Wilson (2001, see especially Chapter 1) should provide a rewarding read.

While these texts do contain some practical advice, other introductions to corpus analysis have a more hands-on focus. McEnery *et al*. (2006) is the only book that we are aware of which provides a 'how-to' approach to using a wide range of corpus search software. By contrast, Hoffmann *et al*. (2008) build their introduction to corpus linguistics around a single tool, BNCweb. Adolphs (2006) has yet a different emphasis, considering the analysis of *texts* as well as corpora

via the methods of corpus linguistics. Finally, Anderson and Corbett (2009) present an introduction to corpus methods using a range of online analysis tools, a kind of software which we will discuss in detail in section 2.5.4.

Of general interest are the various handbooks and readers for corpus linguistics that have been published. Lüdeling and Kytö (2008) and O'Keefe and McCarthy (2010) are two recent handbooks containing a very wide range of helpful readings in corpus linguistics. Both are, however, somewhat expensive and are probably best sought out via a library. More accessible in price is the reader edited by Sampson and McCarthy (2004). This contains a series of 'classic' papers covering a wide range of topics in corpus linguistics.

For those readers interested in the monitor corpus approach, Sinclair (1991), while now somewhat difficult to buy, is available in many libraries. It is a concise introduction not only to the ideas underlying the monitor corpus, but also to many of Sinclair's other thoughts on language. For some reading suggestions on the Web as Corpus approach specifically, please see the further readings section in Chapter 3.

It is harder to make suggestions for readings on non-English corpus linguistics. While there is an increasing amount of research using corpora of other languages, the main textbooks in the field generally remain engaged with English. For this reason, the primary literature – as found in edited collections such as Johansson (2007) and journals such as *Corpora*, *Corpus Linguistics and Linguistic Theory*, and the *International Journal of Corpus Linguistics* – currently represents the best source of material related to non-English corpus linguistics.

Practical activities

As explained in the foreword, we have designed the exercises in this book to be completed with *any* concordancer and with whatever corpus data you have available. The practical exercises for Chapter 1 are a set of very general tasks that should help you find your way around your concordancer if you are not entirely familiar with it. Either using the 'help' file of the software, or else simply by trial and error, try to find out the following things about your concordancer – all of which you will need to know for exercises later in this book.

(A1-1) Firstly, investigate the basic set-up procedures of your software.

- How do you load a corpus into your concordance tool?
- How do you change to a different corpus?
- Does the entire corpus have to be in a single text file, or can your concordancer handle a corpus consisting of many files?
- Does your concordancer need the texts to be in a particular format, or is simple plain text OK?

(A1-2) Next, look at how the concordancing function works.

- How do you search for a particular word?

- Can you search for annotations such as part-of-speech tags, lemmata or semantic tags – assuming, of course, that they are present in your corpus?
- Are searches case-sensitive (treat <A> and <a> differently) or case-insensitive (treat them the same)? Can you change this behaviour?
- Can you *thin* concordances, i.e. reduce the number of results that are displayed?
- How do you save or export a concordance for later reference?

(A1-3) Finally, work out what the statistical capabilities of your concordancer are.

- How can you get a frequency list (of words or tags) in your concordancer?
- Can you get basic corpus summary statistics – such as total number of words (tokens), type–token ratio and so on?
- Can you produce tables of collocation statistics from a concordance?
- Is there a keywords function? If so, how does it work? Can it be adjusted to analyse key tags?
- Can you get a frequency list of *n-grams* (also known as *clusters* or *multi-word units*)?
- How do you save or export these statistical results?

Questions for discussion

(Q1-1) Look at the breakdown of genres within the (hypothetical, non-existent!) corpus of modern British English described in Table 1.2. Is it balanced? Is it representative? Can these claims be made for any corpus sampling frame in an absolute sense, or must they always be qualified?

Table 1.2 *A hypothetical corpus*

Type of text	Number of words
Press (news reports)	7,500,000
Press (opinion columns)	5,000,000
Press (sports news)	5,000,000
Press (culture news and reviews)	5,000,000
Published fiction (books and short stories)	3,500,000
Unpublished fiction (gathered from the Internet)	1,500,000
General non-fiction books	4,000,000
Academic journals (humanities)	500,000
Academic journals (sciences)	500,000
Television programme transcripts (talk shows)	750,000
Television programme transcripts (news broadcasts)	750,000

(Q1-2) Have a look at three or four research papers from the recent primary literature on corpus linguistics – if you can't think what to look at, we suggest any of the following: Culpeper (2009), Calude (2008), Chung (2008), Diani (2008), Hunston (2007), Oakes and Farrow (2007), Inaki and Okita (2006), Biber and Jones (2005), McIntyre *et al*. (2004), Hardie and McEnery (2003), Berglund (2000); links to these papers are available on this book's companion website.

Think about each study's approach to corpus linguistics. Where does it stand, in terms of the different criteria we introduced in this chapter? Remember, you are considering:

- The mode of communication of the corpus that the study uses;
- Whether it is (so-called) 'corpus-based' or 'corpus-driven' in its approach;
- Whether it uses a monitor corpus, a sample corpus or an opportunistic corpus;
- Whether it uses corpus annotations or not;
- Whether it complies with the principle of total accountability or not;
- Whether the corpus data is monolingual or multilingual.

(Q1-1) Imagine a situation where a study has been published that is generally agreed to mark a major advance in corpus linguistics. However, three years later, another study attempts to replicate the analysis and fails – in fact, it gets contradictory results. But the attempted replication was based on a different corpus with a different sampling frame, and a different set of computer programs was used to do the analysis. Obviously, these factors may have had an effect on the results.

How serious a problem would this situation be for the claims of the original study? For example, should researchers avoid any work that relies on its results, pending further replication studies? How often do we need to replicate a contested result before we can accept it as correct? How should we decide to apportion our efforts between replicating existing results versus establishing new results?

2 Accessing and analysing corpus data

2.1 Introduction

The role of corpus data in linguistics has waxed and waned over time. Prior to the mid-twentieth century, data in linguistics was a mix of observed data and invented examples. There are some examples of linguists relying almost exclusively on observed language data in this period. Studies in field linguistics in the North American tradition (e.g. Boas 1940) often proceeded on the basis of analysing bodies of observed and duly recorded language data. Similarly, studies of child language acquisition often proceeded on the basis of the detailed observation and analysis of the utterances of individual children (e.g. Stern and Stern 1907) or else were based on large-scale studies of the observed utterances of many children (Templin 1957). From the mid-twentieth century, the impact of Chomsky's views on data in linguistics promoted introspection as the main source of data in linguistics at the expense of observed data. Chomsky (interviewed by Andor 2004: 97) clearly disfavours the type of observed evidence that corpora consist of:

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this. But maybe they're wrong. Maybe the sciences should just collect lots and lots of data and try to develop the results from them. Well if someone wants to try that, fine. They're not going to get much support in the chemistry or physics or biology department. But if they feel like trying it, well, it's a free country, try that. We'll judge it by the results that come out.

The impact of Chomsky's ideas was a matter of degree rather than absolute. Linguists did not abandon observed data entirely – indeed, even linguists working broadly in a Chomskyan tradition would at times use what might reasonably be described as small corpora to support their claims. For example, in the period from 1980 to 1999, most of the major linguistics journals carried articles which were to all intents and purposes corpus-based, though often not self-consciously

so. *Language* carried nineteen[1] such articles, *The Journal of Linguistics* seven,[2] and *Linguistic Inquiry* four.[3] But even so there is little doubt that introspection became the dominant, indeed for some the only permissible, source of data in linguistics in the latter half of the twentieth century. However, after 1980, the use of corpus data in linguistics was substantially rehabilitated, to the degree that in the twenty-first century, using corpus data is no longer viewed as unorthodox and inadmissible. For an increasing number of linguists, corpus data plays a central role in their research. This is precisely because they have done what Chomsky suggested – they have not judged corpus linguistics on the basis of an abstract philosophical argument but rather have relied on the results the corpus has produced. Corpora have been shown to be highly useful in a range of areas of linguistics, providing insights in areas as diverse as contrastive linguistics (Johansson 2007), discourse analysis (Aijmer and Stenström 2004; Baker 2006), language learning (Chuang and Nesi 2006; Aijmer 2009), semantics (Ensslin and Johnson 2006), sociolinguistics (Gabrielatos *et al*. 2010) and theoretical linguistics (Wong 2006; Xiao and McEnery 2004b). As a source of data for language description, they have been of significant help to lexicographers (Hanks 2009) and grammarians (see sections 4.2, 4.3, 4.6, 4.7). This list is, of course, illustrative – it is now, in fact, difficult to find an area of linguistics where a corpus approach has *not* been taken fruitfully.

The ubiquitous use of corpus data is actually not surprising, given a correct understanding of the analogy between linguistics and the physical sciences. Chomsky's view is at best somewhat naïve and at worst deliberately misleading. Contrary to his assertion, there are entire fields of natural science based not on laboratory experiments but on the collection, and subsequent analysis, of large amounts of observational data: astronomy, geology and palaeontology, to name but three. Even theoretical physics is reliant on real-world observation to confirm or deny its proposals. To give one famous example, it was precise observations of the orbit of the planet Mercury, and of the deflection of starlight by the gravity of the sun, that confirmed the accuracy of Einstein's general theory of relativity, one of the cornerstones of modern physics.[4]

So, unless we are willing to discard as invalid a large part of modern science, Chomsky's argument against corpus linguistics collapses. Just as observation of the universe through astronomy can help to prove the hypotheses of physicists such as Einstein, so observation of language through corpora can help linguists to understand language. But we must not confuse corpus data with language itself. Corpora allow us to *observe* language, but they are not language itself. Furthermore, we do not claim that corpora are the only tool that linguists should use to explore language – introspection and other data collection methods do have their role to play in linguistics. Indeed, without some ability to introspect, it is doubtful whether a linguist could ever formulate a question to ask of a corpus. Nonetheless, there is little doubt that, for the majority of researchers in linguistics, corpora are an indispensable source of evidence, and the tools that extract data from corpora are a substantial and transformative source of support. The utility of

corpus data, and of using it alongside other data, is well summarised by Fillmore (1992: 35):

> I have two observations to make. The first is that I don't think there can be any corpora, however large, that contain all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had the chance to examine, however small, has taught me facts that I couldn't imagine finding out in any other way . . .

Fillmore's sensible methodological pluralism is now widespread, including among linguists who might previously have espoused a much more straightforwardly hostile position towards corpus linguistics. For example, Wasow (2002: 163), a theoretical linguist, observes:

> While data from corpora and other naturalistic sources are different in kind from the results of controlled experiments (including introspective judgment data), they can be extremely useful. It is true that they may contain performance errors, but there is no direct access to competence; hence, any source of data for theoretical linguistics may contain performance errors. And given the abundance of usage data at hand, plus the increasingly sophisticated search tools available, there is no good excuse for failing to test theoretical work against corpora.

2.2 Are corpora the answer to all research questions in linguistics?

If we consider the range of research questions that a corpus on its own allows us to address, we can imagine it as covering a subset of all the research questions that a linguist might ask. That subset will overlap with the subset of questions that a linguist can ask *without* a corpus, but it is almost certainly greater in size than that set. This is because *with* corpus data, we can take new approaches to a number of areas, including grammatical description (see Chapter 4) and even linguistic theory (see Chapters 6 and 7). Moreover, the set of questions that may be readily addressed using a corpus grows significantly as suitable tools for interrogating that data become available. It grows larger still if the corpus data we have at our disposal is suitably annotated with linguistic analyses, in such a way that linguistically motivated queries can be undertaken rapidly and accurately. For example, without a corpus, we could certainly examine, and seek to describe, the use of non-finite verbs in English – many scholars have, for instance O'Dwyer (2006: 58–9). However, the number of examples of non-finite verbs that we could base our investigation on would remain relatively small, being limited by the hand-and-eye techniques that we would need to find them. With a suitable corpus of English, we would have at

our disposal many thousands of words with which to conduct our study. Still, in the absence of tools to search the data rapidly, the exploitation of the data would be slow and prone to error – although we would probably be able to study a greater number of examples somewhat more effectively than we would without the corpus. A search tool that can quickly extract examples of particular words from the corpus would greatly improve the accuracy of our searches (though spelling errors in the data itself might prevent the searches from being wholly error free). Furthermore, the time taken by the searches would fall dramatically. And crucially, corpora allow access to reliable information regarding *frequency*. In the absence of corpus data, even trained linguists find it very difficult to come up with estimates of frequency in language that are reliable (Alderson 2007).

Being able to search for, and extract frequencies of, different wordforms or phrases gets us a long way. But it does not give us all the tools we need for every sort of research question. We can search for *walking*, *having walked* or *to walk*, but we cannot search for 'every non-finite verb'. Searching for each non-finite form of every verb in English would take a very long time indeed. Quantifying the relative frequency of, say, nouns and adverbs in English would take even longer – to the extent that an investigation of these features based on corpus data is effectively impractical if we use searches based on wordform alone. However, if our corpus already has annotations that show the part of speech of each word in the corpus, then, armed with a search tool that understands these annotations, we are able to fashion a query to extract the information we want – the frequency of nouns, or a concordance of all non-finite verbs – rapidly and reliably. So the combination of corpus, search tool and corpus annotation make it possible to explore research questions that would be almost unimaginable otherwise. As studies such as Leech (2004a) and Mukherjee (2005) show, even with something as apparently mundane as a non-finite verb form, a large-scale investigation can reveal features of its use that have escaped linguists who have used intuition or small numbers of examples alone. Indeed Leech (2004a) is a telling example of this. Between its original publication in 1971 and its third edition in 2004, the increasing availability of corpora and tools to search them has led to parts of the work, especially the chapter on modal verbs, being revised substantially. Two good examples are Leech's (2004a) identification of emergent modal constructions such as *need to* and *had better* and his discussion of *would* as pure hypothesis in expressions such as *I would think* and *one would expect*. In both cases it was the data available to Leech which led to his identification of the growing salience of these features in English usage, and hence to the modification of his earlier work. It is difficult to see how such observations could be made reliably on the basis of any other sort of evidence.[5]

Given that, as discussed in Chapter 1, an ever-increasing amount of corpus material is available for an ever-greater range of languages, the remainder of this chapter will look at a number of issues that impact upon the utility of that data. In doing so we will look at two issues mentioned already – corpus annotation and

Table 2.1 *Metadata stored about two speakers, June and Jonathan, in BNC file KCT*

Name:	June	Jonathan
Sex:	Female	Male
Age:	35–44	0–14
Social class:	C2	C2
Education:	n/a	n/a
First language:	n/a	n/a
Dialect/Accent:	East Anglian	East Anglian
Age:	40	10
Occupation:	dinner lady (pt)	student (state primary)
Role:	self	son

corpus analysis tools – and consider, finally, the statistics commonly used to aid the interpretation of corpus data.

2.3 Corpus annotation

2.3.1 Metadata, markup and annotation

Corpora typically contain within them three types of information that may aid in the investigation of the data in the corpus: metadata, textual markup and linguistic annotation. *Metadata* is information that tells you something about the text itself – for example, in the case of written material, the metadata may tell you who wrote it, when it was published, and what language it is written in. The metadata can be encoded in the corpus text, or held in a separate document or database. *Textual markup* encodes information within the text other than the actual words. For example, in a printed written text, textual markup would typically be used to represent the formatting of the text – such as where italics start and end. In transcribed spoken corpora, the information conveyed by the metadata and textual markup may be very important to the analysis of the transcript. The metadata would typically identify the speakers in the text and give some useful background information on each of them, such as their age and sex. Textual markup would then be used to indicate when each speaker starts to speak and when they finish. Consider the examples in Table 2.1 and Figure 2.1. Table 2.1 gives an extract of the metadata relating to a file in the BNC (file KCT). Using information such as that in Table 2.1, it is possible to limit the searches of the BNC in a way which allows a linguistically motivated question to be posed – for example, to extract all examples of the word *surely* as spoken by females aged between 35 and 44.

Figure 2.1 shows an extract of the orthographic transcription contained within the file itself. The BNC is marked up using a specific convention for encoding

Jonathan	357	It's poached
June		<unclear>
Jonathan	358	Egg in the batter.
June	359	Egg in the batter?
June	360	You <-|-> mean erm <-|->
Jonathan	361	<-|-> cooked <-|->
June	362	scotch egg?
Jonathan	363	Yeah, that's it.
June	364	Well you wouldn't like that surely, cos you don't like sausages.
Jonathan	365	They're not sausages.

Figure 2.1 *A conversation between June and Jonathan (BNC file KCT, utterances 357–365).*

such information reliably – the *eXtensible Markup Language* or XML. This is a standard which is used widely nowadays, not merely for corpus files but also, for example, to transfer webpages and word-processor documents reliably from one machine to another. Figure 2.1 does not show the actual XML codes in the BNC file, which are realised as tags delimited by angle-brackets <like> <this>; rather, the text is formatted in a readable way that visualises the structure of the underlying markup.

Together, where they have been introduced into a corpus, metadata and textual markup allow a range of research questions to be addressed. However, we can go beyond merely recording features of a corpus text such as where italics, or the speech of a certain speaker, begin and end. We can also encode linguistic information within a corpus text in such a way that we can systematically and accurately recover that analysis later; when this is done, the corpus is said to be *analytically* or *linguistically annotated*. Annotation typically uses the same encoding conventions as textual markup; for instance, the angle-bracket tags of XML can easily be used to indicate where a noun phrase begins and ends, with a tag for the start (<np>) and the end (</np>) of a noun phrase:

<np>The cat</np> sat on <np>the mat</np>.

How is such linguistic annotation introduced into a corpus and what are the limits on what may be annotated? There are three approaches to linguistic annotation – purely automatic annotation, automated annotation followed by manual correction and purely manual annotation. None of these approaches is currently error free. Automatic annotation of parts of speech in English, for example, can be undertaken to a high degree of accuracy; Garside and Smith (1997) report accuracy of 97 per cent or more. This means, however, that there are still residual errors. Similarly, with manual processes, it is not possible to guarantee that no errors will be made – no human analyst is perfect. Obviously automatic annotation is desirable where its results are accurate enough – it allows new corpora to be rapidly and cheaply annotated. However, it is not currently possible to reliably undertake automated corpus annotation for all types of linguistic analysis.

Nonetheless, a wide range of annotations have been applied automatically to English text, by analysis software (also called *taggers*) such as:

- constituency parsers such as Fidditch (Hindle 1983), used in the production of the Penn Treebank;
- dependency parsers such as the Constraint Grammar system (Karlsson *et al*. 1995);
- part-of-speech taggers such as CLAWS (Garside *et al*. 1987), used to annotate the BNC;
- semantic taggers such as USAS (Rayson *et al*. 2004), which has been used to annotate a number of corpora (see, e.g., Maclagan *et al*. 2008);
- lemmatisers or morphological stemmers, which are often found as built-in subsystems of many parsers and taggers.

Of course, these automatable analyses may still be applied by hand, or at least manually corrected; and there may often be value in doing so, for instance to create a small, *gold standard* dataset to use as a benchmark for measuring tagger performance. The Manually Annotated Sub-Corpus (MASC; see Ide *et al*. 2010) of the American National Corpus, for example, contains multiple layers of manually inserted or manually checked annotation and is clearly suitable for this purpose.

Other forms of annotation have been applied manually, such as pragmatic annotations (see, e.g., Archer 2005). The range of annotations developed and applied is similar for a number of major languages, such as Chinese, French or German. For a host of other languages, however, the availability of programs which can undertake automatic annotation is patchy at best. For example, there is no publicly available part-of-speech tagger for the Bantu language Kinyarwanda that we are aware of. For such languages, either we must develop an automatic system, perhaps using one that can be retrained to work on a different language, or else we must undertake the analysis by hand.

When a corpus includes linguistic annotations, it is important to note what can and cannot be said about those annotations. Firstly, and perhaps most importantly, we cannot say that the corpus contains *new* information. It clearly does not. What a linguistic analysis of this sort does is to make explicit information that is there implicitly in the data. In other words, identifying a word as a noun does not mean that we transform it into a noun in so doing. In corpus annotation we engage in a process of labelling, not creation or transformation. To that extent, we can say that the corpus is *enriched*, from the point of view of a program or user, but we cannot say that the corpus has had new information added to it.

2.3.2 Consistency of annotation

An important point to make, and this point is vexatious, is that we cannot necessarily say that the analyses encoded in the corpus are *consistent*. If the analysis is manual, and if the annotations were undertaken by a linguist

or linguists working to an agreed set of guidelines for applying the annotation, then we can be much more confident in the consistency of the analysis, which is sometimes measured experimentally using statistics which indicate the degree of inter-annotator agreement (Marcus *et al*. 1993; Voutilainen and Järvinen 1995; Baker 1997). So, for example, the SUSANNE corpus (Sampson 1995) was annotated with part-of-speech and constituent structure analysis following a scheme devised by Sampson. This scheme was based upon his earlier experiences annotating the LOB corpus. The annotations were undertaken according to a strict set of guidelines, subsequently published by Sampson (1995). In this case we have a corpus which exhibits a very high degree of consistency, and we have the means to check that consistency by going back to the published guidelines. However, SUSANNE probably represents a high-water mark for human consistency in manual corpus annotation. In fact, it is inevitable that, from time to time, manual annotations will be inconsistent to some degree. Quite apart from considerations of human error, this is due to a property of all linguistic analyses, namely that an analysis typically represents one choice among a variety of plausible analyses. In the phrase *his future bride*, what part of speech is *future*? It may plausibly be considered a noun or an adjective in this case. To some extent the choice between the two options, while it might conceivably be data- or theory-driven, can be seen as arbitrary – that is, choosing either interpretation is fine, as long as this choice is made consistently. But human analysts are typically not very good at being 100 per cent consistent in such decisions.

Computer programs are rather better at being consistent than human beings – given the same input, a program should theoretically always produce the same output (unless any of the procedures it applies involves randomness, but even then the randomness will be introduced in a consistent way). But even an automated tagger may be inconsistent *over time*. The programs used to annotate texts are from time to time updated or changed in some way. For example, the programmer may change their mind about how they prefer to analyse *future* in *his future bride*, and alter the program accordingly. Even if no such changes of interpretation are made, the output of the tagger will change as the algorithms it uses, or the linguistic knowledge base within it, are extended and improved over time. So inconsistency may be inevitable in corpus annotation, whether manual or automatic. But this cannot be seen as a major objection to the practice of corpus annotation because, as we have seen, inconsistency is inevitable in linguistic analysis in general. The virtue of corpus annotation is that, because the choice of the analyst (or annotation software) is explicitly present in the text, any inconsistency is clear and open to scrutiny, and any necessary allowances can be made by users of the data. Without corpus annotation, the potential for inconsistency in analysis still exists, but the prospect of being able to discover it in the work of an analyst is at best remote. This advantage of corpus annotation becomes important when we consider the replicability of our research (see section 1.6.1). A finding based on analyses explicitly available as annotations in a text is fundamentally more

easily replicable than a finding based on analyses of which other researchers do not have a record.

Inconsistency exists at another level in corpus annotation – the consistency of annotations between corpora. Given a scheme for part-of-speech tagging, a team of researchers might work to develop a corpus annotated consistently using that scheme. But such a project would use only one of potentially many schemes for introducing part-of-speech information into corpus data. This is not necessarily a problem – one scheme may simply be more detailed than another. But schemes of analysis can vary in other ways. For instance, linguists may have different views as to what parts of speech exist in a specific language. Even looking at just one language, the same analysts may change their views on what annotation scheme they should use. Over the years, seven different part-of-speech annotation schemes have been developed and applied by the team behind the CLAWS tagger (Garside *et al*. 1987). This variability in annotation schemes can become a problem for users trying to use multiple annotated corpora at once. If the annotation scheme is not consistent across the corpora, then, at best, a translation from one scheme to the other will be needed; at worst, such a translation may not be possible. This problem, though a subject of research for some time (Atwell *et al*. 1994), has come to the forefront for users of annotated corpus data, notably computational linguists (see Chapter 9); and much effort has been expended to develop ways of using corpora annotated with different annotation schemes (see Meyers 2009 for a detailed discussion). Early indications are that the problem is not insurmountable, though, as Meyers (2009: 122) notes, 'a greater degree of coordination among annotation research groups would vastly improve the utility and accuracy of annotation'. This has been tried to some extent in the past, with the Expert Advisory Group on Language Engineering Standards (EAGLES), which worked in the 1990s, being the most prominent example to date.[6] However, no matter how admirable the goal, the ability of linguists to quite legitimately disagree when they are working out schemes for annotating language data means that the labours of researchers like Meyers are likely to be difficult at best and Sisyphean at worst.

Automated corpus annotation is currently conceived of as a distinct type of corpus processing, treated independently of tools for searching a corpus. The overall procedure thus typically has two steps. Firstly, the untagged text of a corpus is loaded into the annotation tool and tagged, producing a version of the text with tags encoded into it using XML or some other formalism. Then, this annotated corpus is loaded into a separate search tool so the researcher can enter their searches and get back the results. These two steps are usually implemented separately, for mainly practical reasons. As we will explain in section 2.5, corpus search tools have over the years developed to become very user-friendly. By contrast, corpus annotation programs – while widely available – typically require so much advanced computer expertise to install and use that they are, effectively, not accessible to most linguists.[7] Moreover, annotating a text is typically a much

longer process than searching it – while nowadays search software can often produce results at a speed that appears pretty much instantaneous to a human being, annotation software is computationally intensive and can take hours or longer to run. So frequently a corpus will be tagged 'once-and-for-all', often by its builders, and the tagged text passed on to users who run searches on it.

There is no reason why tagging and searching have to be separate in this way, however. Let us consider a search for a part-of-speech category such as *noun*. To accomplish this with automatic annotation, the computer has to firstly decide which words in the corpus are nouns, and secondly present all the words that it has identified as nouns. It is entirely possible for these procedures to be carried out at the same time – in which case the intermediate step, where a corpus with annotations encoded into it is saved to disk and then transferred to a different tool, becomes completely unnecessary. There already exist tools which bring together annotation and search under a single roof, such as GATE,[8] Wmatrix (Rayson 2008), and SketchEngine (these latter two tools will be discussed in detail later on). In these cases, however, the tagging and searching systems are still independent, behind the scenes. More notably, the Nooj system (Koeva *et al*. 2007) is a complex, multilingual annotator-cum-concordancer in which the procedure for annotating some linguistic feature and the procedure for searching for it are one and the same.

We may expect that, in the future, joint annotation–search tools of both these kinds will become more common. In many ways this development is much to be desired, as it will open up automated annotation to users of corpus data who could not hope to manage the technicalities of much tagging software. It will also free users from the strictures of the annotation applied to a corpus by its builders, allowing them to choose what annotations they want to apply. There is, however, also a downside to this kind of on-the-fly annotation. Firstly, it removes the possibility of running manual checks on the output of an automated tagger before using it, whereas this is very easy when annotation and search are firmly separated. While this is not an unimportant consideration, it is in practice rare for such checks to be carried out in studies where the annotation is ad hoc, rather than part of the preparation for releasing a dataset publicly. In most such studies, the published error rates for a tagger are simply accepted and allowed for, so in these cases nothing would be lost by joint annotation–search. Secondly, and more seriously, if no annotated version of the corpus is actually created on disk and archived, then the advantages of corpus annotation with regard to replicability that we discussed above are undermined. Tools such as Wmatrix, where a copy of the text with encoded annotations *is* created in the background and is available to the user, may therefore be preferable to combined annotation–search tools that do not have this feature.

Combined annotation–search tools may be the direction of the future, but today we must deal in large part with tools that are designed for searching alone. A range of such tools is vitally necessary to explore a corpus in service of our research question. A well-annotated corpus may be of inestimable value, but

把 他 打 得 **晕头转向** ， 他 不得已 **报**	**了**	**师范** 院校 ， 尽管 **录取** 他 的 **师范大学** 依
觉得 自己 已 从 宝塔 **顶** 上 跌落 下来	**了**	。 入学 后 ， 肖 立 从不 佩戴 校徽 ， 未来
不 知道 ， 儿子 的 枕 下 已 放 **满**	**了**	武侠小 **说** 、 黄色 **书刊** 。 直到 大学 四 年 **级**
， 四 门 功 **课** 不 及格 的 **现实** 一下 **压垮**	**了**	肖 立 。 茫然 中 ， 他 想到 死 。 在 死
欢乐 ， 于是 ， 他 魔鬼 般 地 扑 向	**了**	**邻居** 的 一个 女孩子 ... ... 他 在 惊慌 中
子 ... ... 他 在 惊慌 中 残忍 地 打 昏	**了**	女 孩儿 ， 自己 跌跌撞撞 地 逃 出 了 小
了 女 孩儿 ， 自己 跌跌撞撞 地 逃 出	**了**	小屋 ... ... " 当 我 打 那 女孩 儿 **时** ， 我
可是 ， 她 没 昏 ... ... 我 慌	**了**	， 束手无策 ... ... 那 一会儿 ， 我 **脑子** 里
第二 次 **谈话** **时** ， 我 和 他 一起 分析	**了**	他 犯罪 的 原因 。 他 **说** ： " 如果 可能

Figure 2.2 *An extract of a sample concordance of the particle 了 from the LCMC.*

realising that value is dependent on being able to search it in a way that is reliable and linguistically motivated. So what tools of this kind have been developed, and what kinds of analyses do they enable?

2.4 Introducing concordances

Undoubtedly the single most important tool available to the corpus linguist is the concordancer. A concordancer allows us to search a corpus and retrieve from it a specific sequence of characters of any length – perhaps a word, part of a word, or a phrase. This is then displayed, typically in one-example-per-line format, as an output where the context before and after each example can be clearly seen. Figure 2.2 shows a concordance for the Chinese word *le* (了), taken from the Lancaster Corpus of Mandarin Chinese (the LCMC; McEnery and Xiao 2004a) using the CQPweb concordancer (Hardie forthcoming).[9]

We have avoided saying that a concordance shows *words* in their context – though this procedure is often called *key word in context* (KWIC) concordancing, KWIC need not be limited just to showing whole words. For example, in English we might want to produce a concordance of a common suffix in order to explore its context of use – Figure 2.3 shows an example of this, a concordance of words ending in the nominalising suffix *-ness*.

Likewise, if we have an interest in idioms, we may wish to view a concordance of a multi-word expression. Importantly, if there is some form of linguistic annotation embedded within the corpus data, we may wish also to search not for a linguistic form at all, but rather for one of the annotation mnemonics – so, for example, a search for words marked with the part-of-speech tag *u* will return all of the 75,273 auxiliaries within the LCMC.

Exeter : the capital city of	**blandness**	ca n't break the chain TWO years ago the
after Yahoo warned it was seeing	**weakness**	in two of its biggest advertising segments .
advertising in the face of economic	**weakness**	. ''The warning , which was similar to
in the afternoon following news of more	**weakness**	in the US housing market , with housing
became a lifeline for Ross when his	**illness**	kept him confined to hospital and his home
Olympic bid . She was briefly married to	**fitness**	guru John Crisp . Her second marriage to
and are far less secure . " Ironically , this	**hyper-attentive-ness**	is actually having a damaging effect on our
thing is absolutely barmy . " And the	**madness**	continues with the goalposts constantly
notes led to a lifetime of passion and	**happiness**	. She was just 17 and he was 18 when they
so much , just to feel your warmth and	**warmness**	around me (lovely , lovely thought) . I 'm
Capt Alfred Bland MY only and eternal	**blessedness**	, I am never utterly miserable , not even
will and giving up a chance of supreme	**happiness**	with you . What more can anyone ask ?
made me during those days-just you , my	**sweetness**	. MY darling ... the nights here are weird .
range hens under threat as devastating	**illness**	nears Britain POULTRY farmers in
and has submitted it for inclusion in the	**Guinness**	Book of Records . 'Unfortunately , the

Figure 2.3 *An extract of a concordance of words ending in* -ness *from BE06 (Baker 2009).*

Corpus search tools are vital if the range of research questions we can address using a corpus is to be significantly expanded. But while concordancers, and other tools for exploring a corpus, are powerful aids to the linguist, they also, crucially, limit and define what we can do with a corpus. On the other hand, it is for practical reasons impossible to avoid using these tools. The analysis of very large corpora without computer processing can best be regarded as what Abercrombie (1965) referred to as a *pseudo-procedure* – a method in linguistics that might yield very interesting results and be useful in principle but which is, for all practical purposes, impossible. It is simply too time-consuming. In fact, fundamentally, the corpus-based approach to language cannot do without powerful searching software. Conversely, concordancers of suitable power remove the pseudo-procedure argument for that set of research questions which a concordancer, working in concert with a specific corpus, can address. For example, consider the LCMC. This is a million words of Mandarin Chinese from the early 1990s which has been annotated with part-of-speech information. The corpus was collected according to the LOB sampling frame. There is a range of questions which can be addressed using this corpus; equally there is a set of questions that cannot be answered readily, or even at all, using the corpus. Some of these questions are excluded by virtue of the composition of the corpus; it contains no nineteenth-century texts, so it cannot be used to explore the Mandarin of that period. However, some questions are excluded by virtue of the available annotation. While we can use the part-of-speech data in the corpus to explore various research questions (many relating to Mandarin grammar), there are, equally, questions that we cannot answer because the annotation we would need to perform the necessary queries is not present or

not searchable. For example, there is no annotation of constituent structures in the LCMC. In the absence of an automatic phrase-structure parser, and a tool to search its output, the use of the corpus to study constituent structure in Mandarin may well still be considered a pseudo-procedure. Of course, we might approach constituent structure indirectly by means of searches for words or part-of-speech tags. But it is impossible to *directly* address this research area without the appropriate annotation. This is an important point; the corpus alone solves few (if any) problems for a linguist. Its potential is unlocked by tools that allow linguists to manipulate and interrogate the corpus data in linguistically meaningful ways. The availability of tools that are relevant to specific research questions remains a crucial limiting factor in corpus linguistics. Over time, an increasing number of tools *have* become available, and this has expanded the range of research questions that may be addressed using a corpus. But a range of research questions do still lie beyond the reach of what can be done with a corpus at the present time. We will return to this topic in section 2.5.4.

2.5 A historical overview of corpus analysis tools

When did the process of developing software tools for corpus analysis begin? The honour of doing this may reasonably be claimed by Roberto Busa, who built the first machine-readable corpora and undertook the first automated concordances in 1951. Busa did not, however, invent the concordance; although it was in the realm of the pseudo-procedure for most purposes, some hand-compiled concordances had been available for some key works for a long time. For example, Hugh of St Cher, with the assistance of around five hundred monks, compiled the first concordance of the Latin Vulgate Bible in 1230, providing, for each word, an index of where each instance of it could be found. However, Busa showed that, with a little effort, concordancing could be applied rapidly and effectively to electronic texts. This was a pivotal moment, when concordancing moved from being a labour-intensive task applied to a few texts of particular cultural importance – such as the Bible, the Qur'ān, or the works of Shakespeare – to being a technique that could, in principle, be applied to any text at all. Busa's work led to what we will term *first-generation* concordancers.

2.5.1 First-generation concordancers

First-generation concordancers were typically held on a mainframe computer and used at a single site, such as the CLOC (Reed 1978) concordance package used at the University of Birmingham.[10] Individual research teams would build their own concordance system and use it on the data they had access to locally. The packages typically did no more than provide a KWIC concordance. Any other manipulation of the data was undertaken by separate programs. So, for

example, to build a list of the words appearing in a corpus, a different program would be used, such as that used by Hofland and Johansson (1982) to produce word frequency lists of English. First-generation concordancers typically had great difficulty dealing with anything other than the non-accented characters of the Roman alphabet – where characters carrying diacritics appeared in English texts, for example, they would be replaced by character sequences designated to represent those characters. These character sequences were usually agreed on a site-by-site basis. For example, in LOB, an apostrophe after a vowel indicates an acute accent on the preceding letter – so the word *café* would be encoded as *cafe'*. (More recently, standardised sequences such as *café* in which the XML code *é* replaces the *<é>* character have been used.) This points to a general limiting factor that runs through the history of concordance software. Where an international standard did not exist for some feature of corpus use or storage that had to be dealt with, concordance program writers and corpus builders improvised a way around the issue. Another example of this relates to how markup and annotation were encoded in corpora. Standards such as XML (see section 2.3.1) have now emerged to allow information such as linguistic annotations to be reliably inserted into corpus texts. But no such standards existed to begin with. Accordingly, as with character encoding, ad hoc standards emerged in those centres developing corpus texts and software tools. So, for example, at Lancaster in early corpora such as LOB (Johansson *et al*. 1978) and the Spoken English Corpus (Knowles 1993), an underscore character was used to associate a word with its part of speech. So an instance of the word *dog* acting as a noun in the corpus would appear as *dog_NN1*, where NN1 is the mnemonic for a singular common noun. Tools which manipulated these corpora, such as Roger Garside's XANADU (Fligelstone 1992), were programmed to interpret annotations of this sort. First-generation concordancers were useful in development terms as they clearly showed the need for standards – if corpora were to be passed between users at different sites, then some agreement had to be reached on how this kind of information was to be encoded. They also showed how wasteful it was for people to reinvent the wheel. It was clearly preferable to write concordancers that could work on a range of machines and a range of corpora. There was a further clear advantage to doing this: replicability (see section 1.6.1) was difficult to achieve in an era when, though corpora were shared, the tools for manipulating them were not. While it might be hoped that results would be broadly replicable between sites, without both the corpora and the tools on which a study was based being made widely available, it was difficult to test the replicability of any of the results produced at a site without visiting that site and getting permission to work there.

First-generation concordancers also clearly demonstrated that, rather than having stand-alone tools for functions like generating frequency lists, it was preferable to write software which bundled many tools together into a single package which, while it might still be called a concordancer, would allow the data to be manipulated in as wide a range of ways as possible. The widespread

availability of personal computers from the 1980s onwards removed a major barrier to solving these problems. A problem with mainframe computers was that programs written to run on one mainframe would often not run on a mainframe of another type. While some effort might make the program work, that effort might be so great that just starting from scratch and writing your own KWIC concordancer might seem preferable.

2.5.2 Second-generation concordancers

Second-generation concordancers were enabled by the spread of machines of one type in particular across the planet – IBM-compatible PCs. It became possible to write and distribute a concordancer for that platform with a high likelihood that the program would be usable straight away on the recipient's machine. This was important in two ways. Firstly, it meant that a lot of effort that had gone into reinventing the wheel could now be directed towards producing better tools. Secondly, it had a democratising effect. Up to this point corpus linguists typically needed to work in a team which included a computer scientist who was prepared to do whatever programming was needed, either to produce tools or to make somebody else's tools work on the local mainframe. With PC-based concordancing, any linguist who was able to switch on and use a PC could use corpora. The effect was fairly immediate. Corpus linguistics boomed from the late 1980s onwards. The increasing availability of tools such as the Kaye concordancer (Kaye 1990), the Longman Mini-Concordancer (Chandler 1989) and Micro-OCP (Hockey 1988) enabled this boom. Second-generation concordancers, however, inherited many of the problems of the first generation. They provided very few tools other than KWIC concordancing. They could typically sort alphabetically the left and right context of the word searched for, produce a wordlist, and calculate some basic descriptive statistics about the corpus (word count and type–token ratio, for example). They also had the same issues with character representation and corpus formatting as the first-generation concordancers – standards had not developed to the point where generally agreed formats could be relied upon.

Partly as a result of this, the second-generation concordancers were, arguably, worse than the first generation. The second-generation concordancers generally made no assumptions about a corpus being marked up in a particular way. So, for example, when using the Longman Mini-Concordancer with a corpus marked up in the Lancaster convention of *word_TAG*, the user had to be aware of and understand that convention, and develop searches accordingly. For instance, a search for *dog* would find nothing if all instances in the corpus appeared as *dog_NN1*! The computer just interpreted the corpus as a very long stream of undifferentiated text – it did not 'know' what parts were words and what parts were part-of-speech mnemonics. By contrast, the earlier bespoke programs written for a specific group were often able to prise the two apart. First-generation concordancers were also superior to the second generation in terms of *scale*, that

is, the size of the corpora they could successfully search through. For instance, because of the limitations of early PCs, the Longman Mini-Concordancer could search through a few tens of thousands of words before it ran out of memory. Meanwhile, at that point in history, tools running on mainframes were able to deal with corpora of a million words or more, albeit slowly at times. However, the energy released by ending the need to reinvent the wheel, and the continuing increase in the power of PCs, led to a third generation of concordancers.

2.5.3 Third-generation concordancers

The third generation of concordance software includes such well-known systems as WordSmith (Scott 1996), MonoConc (Barlow 2000), AntConc (Anthony 2005) and Xaira.[11] These concordancers were able to deal with large datasets on the PC (the 100-million-word BNC is packaged with Xaira). Moreover, they had bundled in with them a wider range of tools than had previously been available in concordance packages, and they gave access to some meaningful statistical analyses which went beyond the merely descriptive. Finally, they effectively supported a range of writing systems.

The last point deserves a little more discussion. Stable and widely accepted standards for encoding information in corpora – notably XML – have developed in recent years, as discussed previously. At the same time, the encoding of the actual characters in a corpus has also standardised around a system called Unicode. Before Unicode it was very difficult for a system developed to support the Roman alphabet to read, appropriately display and effectively analyse corpora in other writing systems. With the advent of Unicode, computers and the programs that run on them have become much more effective at dealing with data in a range of writing systems. Third-generation concordancers have developed to exploit this new capability – for instance WordSmith, as of version 4, can be used to analyse Unicode-compliant corpora. This represented a great breakthrough for corpus linguistics and, if nothing else, has saved a great deal of effort. Prior to Unicode being widely implemented, concordancers had to be developed to cover each writing system separately, leading to the production of concordancers that could only be used with certain scripts such as Arabic (Abbès and Dichy 2008) or Chinese (Luk 1994). Where concordancers were developed that could be used with a range of scripts (Wools 1998), effort and some technical expertise was required by the user to retarget the concordancer at a corpus in a new writing system. Unicode has removed the need to redevelop basic packages to support concordancing across writing systems. In so doing, it has greatly simplified the task of working with a disparate range of languages and has increased the ease with which corpus linguistics can be adopted. As more and more Unicode corpora become available, the usefulness of the Unicode-compliant concordancers will become ever more apparent.

It is notable that the third generation of concordancers are in many ways remarkably similar to each other, especially in terms of their core functionality: concordances, frequency lists, collocations[12] and keyword analysis[13] are the main tools available in each. The ubiquity of these four analysis functions is not surprising. Given a corpus of texts, a computer can do two basic things: count the things in the corpus (a frequency list) and locate all of the examples of a search term and display them (a concordance). It can then derive statistical abstractions on each of these outputs: keywords are a statistical abstraction from frequency lists and collocations are a statistical abstraction from a concordance. Each of these procedures may then be carried out on annotations as well as words (so, we can get a frequency list of part-of-speech tags, or calculate key tags, or calculate which tags collocate with a particular search term), although the degree to which tools allow annotations to be referenced independently of words can vary. Beyond these core procedures, no others have yet emerged as widely agreed to be essential for a concordancer. In short, then, the set of functions which analysts are using to explore corpora seems to be relatively stable at present. There are differences in the specialisations of particular tools, of course. Xaira, for instance, does not include any tool for the calculation of keywords, a core function of WordSmith and AntConc; but it does provide powerful support for searching for XML elements (whether markup, annotation or metadata). Tools also differ in the extent to which they support lists of sequences of words in a corpus (*n-grams*). Yet at their core the third-generation packages are, we would argue of necessity, similar.

Given that the tools embedded in these programs were largely available during the first generation of concordancing – albeit to a much smaller audience – should we conclude that corpus exploitation techniques reached such a level of maturity relatively early on that all that is left to do at present is to repackage these tools, and perhaps make them easier to use? Almost certainly not. Three pieces of evidence tell against this conclusion. Firstly, there are tools from the first generation of concordancing that are not generally available in third-generation concordance packages. A good example of this is the technique of *collocational networks* developed by Phillips (1989). This method appears to be useful, and has been used since Phillips developed it, but always in the form of bespoke software which is neither embedded in popular concordance packages nor publicly available (see McEnery 2005: 20–3). Collocational networks provide a telling example of how the inventory of tools developed in the first phase of corpus software development has yet to be transferred in full into the third phase. Another example of the same phenomenon is the multi-dimensional (MD) approach of Biber (1988; see section 5.4). MD is a highly influential approach to the analysis of corpus data; yet although the general thrust of the approach has been roughly duplicated using general corpus searching software (Tribble 1999; Xiao and McEnery 2005), there is no easy-to-use integrated package publicly available that will perform a full MD analysis from beginning to end in a sufficiently user-friendly way to make it accessible to the majority of linguists. Secondly, as we can see from

the current literature, there are techniques that have been developed recently which are not incorporated in contemporary concordancers. A good example of this is collostructions (Stefanowitsch and Gries 2003; see section 7.5.1). These are potentially important to discover in corpora, yet third-generation packages such as WordSmith and AntConc do not directly support users in extracting them. As with MD, with a little ingenuity some existing tools, notably collocation packages, can be used to begin the process of finding collostructions, but this falls short of the necessary full support for discovering them with accuracy, speed and ease. Finally, there are specialised concordancers which, while lacking the range of features available in popular third-generation concordancers such as WordSmith, nonetheless provide tools which are of clear relevance and importance to linguists. Probably the most striking case of this is the ICECUP program (Quinn and Porter 1996) provided to support the analysis of ICE-GB (the British component of the International Corpus of English).[14] ICE-GB has been annotated with syntactic tree structures; it is thus a so-called *treebanked* corpus. In a treebank, the main syntactic constituents of each sentence are annotated; one common system of annotation uses brackets to which a mnemonic tag is added to denote the constituent (S for simple declarative clause, NP for noun phrase, PP for prepositional phrase, VP for verb phrase, etc.). The following example from a treebanked corpus using this format is drawn from Taylor *et al*. (2003: 7):

> ((S (NP Martin Marietta Corp.) was (VP given (NP a $29.9 million Air Force contract (PP for (NP low-altitude navigation and targeting equipment))))).)

Treebank annotation is typically very complex and multilayered – the parsing structures embedded in ICE-GB are considerably more detailed than the example above. The ability to search a treebank rapidly and effectively is clearly a significant advantage, as is the ability to get clear graphical displays of the output, since the underlying bracketed text gets harder and harder to read as the analysis gets more detailed. ICECUP has both these capabilities. As such it is a highly useful and sophisticated piece of software, despite the significant drawback that it can only be used with the style of syntactic annotation employed by the team of linguists who created ICE-GB. Yet this search and display functionality is currently lacking from concordancers in general use. In this case we might be able to address the shortcomings of the general packages by switching out to a specialised program such as ICECUP to carry out relevant searches where necessary. But at best this can be described as a sub-optimal solution, and it is probably better characterised as quite unsatisfactory.

In short, there is plentiful evidence – from tools developed in the past and in use at present – that the range of the existing corpus analysis software could very usefully be extended, either in terms of user-friendliness or function. This is clearly an important issue in corpus linguistics; as we have noted, if the toolset does not expand, then neither will the range of research questions that may reasonably be addressed using a corpus. We might ask *why* general corpus search

tools do not incorporate these kinds of extended analyses. The answer may be different for each type of analysis, but in general it is related to the limited resources of developer time available for the creation of a software tool. The third-generation concordancers were based largely upon the efforts of talented and committed individuals – for example, AntConc was developed by Laurence Anthony and WordSmith was developed by Mike Scott, each working alone. While both were assisted to a degree by enthusiastic users who helped them to debug and expand the functionality of their programs, they laboured on their packages with no external aid, relying instead on their considerable talent and deep commitment to helping others. While this is entirely noble and admirable, there are also hard limits to the amount of time available to individuals working in this way. This being the case, it is unsurprising that they have concentrated on implementing the corpus analyses – concordances, collocations, keywords – that are of the greatest generality and of use to the most researchers. Many of the extended analyses we have discussed are restricted in their application for either practical or theoretical reasons. For example, as noted, ICECUP only works with corpora annotated in the style of parsing used at the Survey of English Usage (SEU) research unit at UCL. It cannot be used to search any other parsed corpus (or, indeed, corpora with other kinds of annotation). But there is a very wide variety of parsing schemes, of which the SEU's system is only one. This explains to a large extent why no ICECUP-like search and display system is found in any general tool: the tool's author would have had to expend a great deal of effort to implement this functionality, for the benefit of a relatively narrow subset of the users – those using corpora with SEU-style syntactic annotation such as ICE. Similarly, collostructions are associated with a particular theoretical approach to language (namely Construction Grammar). So implementing direct support for collostruction searches in a general corpus package would only be of help to users working within theoretical frameworks compatible with Construction Grammar. The limitations of the general software are, then, quite understandable. Nonetheless, it is still deeply regrettable that the basic toolset places these implicit restrictions on the expansion of the field of corpus linguistics.

2.5.4 Fourth-generation concordancers

Considering the need to expand the range of corpus analysis tools that are available, it is a pity to note that the fourth generation of concordancers are strikingly similar, in terms of their functionality, to the third. In fact, fourth-generation concordancers have arisen, not to extend the range of available analyses but to address three entirely different issues: the limited power of desktop PCs, problems arising from non-compatible PC operating systems and legal restrictions on the distribution of corpora. We will discuss the legal issues associated with corpus construction in the next chapter. However, it is often the case that a corpus, once collected, cannot simply be given away to any researcher who wants to use it, because that would violate the original text producers'

copyright in the texts within the corpus. This is a deeply frustrating restriction for the corpus builder; it means that fewer people will be able to take advantage of their hard work, and it also reduces the scope for replicating the results achieved with one particular corpus. In the past, corpus builders would get around this issue by distributing the corpus only to institutions who could officially sign up to a restrictive licence, or simply by not distributing the corpus. However, a preferred solution more recently has been to make the corpus available through a web-based interface. That is, a website is created where users can enter search queries and get back dynamically generated results. The proportion of any given text contained within a concordance line is usually no more than a short chunk of a sentence. This is typically thought to fall within the level of 'fair use' allowed under copyright law, and thus does not violate the rights of the text producer (although the legality of such 'fair use' redistribution has yet to be comprehensively tested). For example, Mark Davies made the BNC available via such a website; the Polish PELCRA corpus and the Hellenic National Corpus are examples of corpora whose *primary* public availability is via a web-search interface.[15] These interfaces were not only motivated by legal considerations, of course. They also allowed corpus builders to make their work available immediately, and via a piece of software (the web browser) that all computer users are already familiar with. By being available across the web, they were instantly available to users on any operating system – in contrast to third-generation tools such as WordSmith, only available on Microsoft Windows, or Xkwic, only available on Unix-like systems. Finally, these web-based systems are typically capable of much faster searches than a PC-based concordancer. The corpora to which they offer access are often very large, on the order of a 100 million words. Searching such a corpus on a desktop PC can be a very lengthy process; while modern PCs are very much more powerful than those that the Longman Mini-Concordancer was developed for, any corpus software designed for a PC is still subject to the limitations of memory and processing power of a given user's computer. But by using the web, the fourth-generation tools have effectively decoupled local processing power from the issue of corpus searching. Searches can be completed much faster than the same searches would be if running on the desktop. Thus, these interfaces combine the advantages of the PC and 'mainframe' approaches – users are given access through their browser to a distant, sometimes more powerful, machine on which the actual corpora and corpus processing tools are held. This is known technically as a client/server model – the PC has on it a 'client' program, mediating between the user and the machine on which the real analysis of the data takes place, the 'server'. The client passes requests from the user to the server, and relays back to the user the results of the server's work.

We consider the use of this client/server model, via the specific medium of the World Wide Web, to be the defining feature of fourth-generation concordance tools. Client/server systems already existed in the third generation. Xaira, for example, consists of a server program and a client program (they often run on the

same machine, but do not have to). Xkwic is another example of this arrangement. But it is the move onto the web that frees the tool from the restrictions of local processing power and makes the user's operating system irrelevant. In web-based concordancers, the only thing that is done locally is the rendering of the webpage that contains the results. Numerous web browsers exist for all operating systems, some commercially developed and some not; most computers now come with at least one pre-installed. So the software on the client end can be taken as read – it is only the software on the server end which needs to be developed. The server software normally consists of a powerful and complex search system, which usually works on a corpus that has been *indexed* in some way. Indexing refers to any processing of the corpus which allows words, phrases or tags to be looked up in a search *without* the program going through the entire corpus in search of them. Indexing is crucial if very large corpora (tens or hundreds of millions of words) are to be searched at an acceptable speed. The details are highly technical and vary for each system. But in general, there are two types of software used at the server end. The first is a database system. If a corpus is loaded into a standard database management program, then concordances can be carried out very efficiently using the normal query language of that database system. Since databases are an important application of information technology in general, extremely powerful database software has been developed over the years. Most databases use a formalism called the Structured Query Language or SQL. Examples of SQL-based web interfaces include the already-mentioned PELCRA reference corpus of Polish and Mark Davies' corpus.byu.edu interface (Davies 2005, 2009a).[16] The other type of software that can be used is a dedicated corpus indexing and querying system. The mostly widely used of these is CQP, the Corpus Query Processor, which is a part of the Open Corpus Workbench (CWB),[17] developed at the Institut für Maschinelle Sprachverarbeitung at the University of Stuttgart (see Christ 1994). CQP is the server program behind the Xkwic concordancer. However, it is now more often used as a back-end to web-based interfaces.[18] The Xaira server program can also be used as a website back-end.

While fourth-generation corpus analysis tools began as websites allowing the searching of specific corpora, they have now been extended beyond this into generalisable systems. Three of these are of particular note. The system developed for the BNC by Mark Davies, mentioned above, has been extended by him to allow access to a wide range of very large corpora via his corpus.byu.edu site. This tool is now effectively non-corpus-specific. It is probably the most powerful, and most widely used, SQL-based corpus analysis tool at the time of writing. Another important system is SketchEngine, which uses a CWB/CQP-compatible program called Manatee as its back-end and allows users to analyse a wide range of corpora for lexical and lexicogrammatical patterns; SketchEngine is particularly useful in support of corpus-based lexicography (Kilgarriff *et al*. 2004). Finally there is BNCweb (Hoffmann *et al*. 2008) and its clone CQPweb (Hardie forthcoming),

which combine an SQL database with a CQP back-end. As with Davies' (2009a) system, the original BNCweb is an interface just to one corpus (the BNC), and its re-engineering as a corpus-independent system, CQPweb, was a later development. But as noted above, these systems have very similar functionality to third-generation software such as WordSmith, though they have made progress by adding support for annotation and for regular expressions,[19] which allow more complex patterns to be retrieved from the corpus.

Fourth-generation corpus analysis tools are even more user-friendly and powerful than the third-generation tools (although they are typically not as useful in the study of very small corpora or individual texts). But they do not noticeably expand the range of analysis tools available within general purpose packages. How can this be addressed? If the market for corpus processing software expands sufficiently, it is not inconceivable that a major software company might develop a corpus processing system. Teams of professional programmers could easily produce a system to surpass what was possible for the pioneering individual developers of the third-generation systems. However, to date no major software company has undertaken such an enterprise. Instead, the pioneers of the third- and fourth-generation concordancers continue to push on. One promising approach is open-source modular concordancers (Anthony 2009). The idea here is that anyone working with corpora could create a tool within the framework of an existing concordance program whose source code is openly available. The new tool is then released to all users of the original system. This reduces the amount of work that has to be done to get the new tool up and running, and means that less reinvention of the wheel is done. If this approach was adopted in the context of fourth-generation concordancers, then a comprehensive solution to the problems facing corpus tools developers and users could well emerge. However, there are certain practical difficulties. Most corpus search and retrieval programs are hugely complicated systems, so complicated that it can be very difficult for anyone other than their primary maintainer to understand their internal workings well enough to begin to program extensions. For instance, while both Xaira and CWB have been open-source software for several years at the time of writing, the overwhelming majority of the development work is done by one or two people in each case. As an alternative to the idea of an open-source, modular approach, it has been suggested that corpus linguists, rather than using general corpus analysis packages, should instead fully embrace computer programming and individually develop their own ad hoc tools to address the tasks that face them. This view is probably most forcefully put by Biber *et al*. (1998: 254) who see the following advantages to this approach:

> concordancing packages are very constrained with respect to the kinds of analyses they can do, the type of output they give, and, in many cases, even the size of the corpus that can be analyzed . . . Computers are capable of much more complex and varied analyses than these packages allow, but to take full advantage of a computer's capability, a researcher needs to know how to write programs.

In particular they see the key advantages of this approach to be that (Biber *et al.* 1998: 256):

- you can do analyses that are not possible with concordancers;
- you can do analyses 'more quickly and more accurately';
- you can tailor the output to fit your own research needs;
- you can analyse a corpus of any size.

Each of these advantages is potentially significant where this approach may be taken. However, there are notable limits to this approach:

1. Not all linguists want to be computer programmers – second-generation concordancers promoted corpus linguistics by providing tools to the significant majority of linguists who may want to use corpora without becoming computer scientists *manqués*.
2. Small teams or individuals working alone can only go so far. By coordinating the work of teams towards a common goal we might reasonably expect greater progress in corpus tool development to be made.
3. By working together on corpus and tools development, the process of the development of corpus encoding standards was accelerated. Without this impetus to develop standards, it is unlikely corpora would be as widely used as they are today, as they would be mired in conflicting formats with tools being developed and redeveloped to deal with those formats.
4. Locally developed tools need to be made available to other researchers so that findings based upon them may be replicated and verified. Where tools are made widely available by developers, the user of the tools is not faced with this demand.

However, Biber *et al.*'s points are not necessarily in conflict with the development of general purpose corpus analysis packages. It is often by teams, especially interdisciplinary teams with expertise in linguistics, programming and statistics, working in the framework that Biber has suggested, that new tools are developed in the first place – only later finding their way into concordance packages. Similarly, in Anthony's vision of a cooperative approach to corpus software development, the work of isolated linguists building tools is integrated into widely available packages; in this way, the efforts of those linguists who wish to undertake computer programming, or who collaborate with computer programmers, could be made available to the benefit of all linguists. In that spirit, it is a good idea indeed for those linguists who can do so to continue to develop new tools for corpus exploitation. There is certainly more support available to them through books such as Mason (2001), Gries (2009a) and Weisser (2009) than there has ever been before. Yet for those who cannot or do not wish to program, the fruitful avenue of cross-disciplinary cooperation is still open.

Before leaving the question of the future development of corpus processing tools, however, it should be remembered that if there are problems in corpus processing to be addressed in the future, dealing with those problems is now possible because some basic problems of the past have been addressed. An accepted standard now exists, in the form of Unicode, for encoding a range of writing systems. Likewise generally accepted standards for formatting information in corpora now exist, such as XML. The utility of annotating linguistic analyses in corpora is widely accepted, albeit not universally. Corpora and corpus processing tools are now available virtually to all, not just to a small number of well-funded groups. All of these very real developments have allowed corpus linguists to look to a future where further tools, of great utility to all languages and the majority of linguists, might be realisable. These may include concordancers which are just as useful when dealing with audio and video recordings as they are when dealing with textual material.[20] Allowing for a clearer flow of information between concordancers and database and spreadsheet programs may also greatly facilitate the ease with which corpus data may be manipulated.

2.6 Statistics in corpus linguistics

Corpora are an unparalleled source of quantitative data for linguists. It is hardly surprising, therefore, that corpus linguists often test or summarise their quantitative findings through statistics. It is possible to use a corpus and, quite legitimately, not engage in statistical analyses. For example, to establish that a particular phoneme *exists* in a given language, all we have to do is find one good, well-attested example of that phoneme occurring. This is an all-or-nothing assessment – finding a thousand examples does not necessarily help us any more than finding one. Of course, *not* finding an example does not necessarily count as evidence of non-existence (though it may count as evidence of rarity).

However, if we want to establish that the same phoneme is *frequent* in that language, we need to employ statistics, because frequency is not an all-or-nothing matter. In this case, we would need some basis of comparison on which to assert that the phoneme is frequent. Frequency cannot be measured in an absolute sense such that *frequent* has an invariable value associated with it (saying flatly 'anything above twenty counts as "frequent"!' would be silly). Typically, *frequent* is a relative judgement. If we have a frequency list of phonemes from a phonemically transcribed corpus, we could say that phoneme X is one of the most frequent phonemes, based on its *relative* position on the list. We might also consider the frequency of phoneme X in this corpus *relative* to its frequency in some other corpus. Straightaway, we need an understanding of how to calculate relative (or *normalised*) frequencies, and we may well also need an understanding of how to

apply a statistical significance test to differences in frequency, to assure ourselves that the results our statistics produce have not occurred by simple coincidence. We will return to both these procedures later on.

Corpus linguistics is not unique in linguistics in appealing frequently to statistical notions and tests. Psycholinguistic experiments, grammatical elicitation tests and survey-based investigations, for example, all commonly involve statistical tests of some sort being carried out. However, frequency data is so regularly produced in corpus analysis that it is rare indeed to see a study in corpus linguistics which does not undertake some form of statistical analysis, even if that analysis is relatively basic and descriptive, for example using percentages to describe the data in some way. To put it another way, any empirically based approach to linguistics which deals with large collections of data points may have cause to employ statistical analysis. Empiricism lies at the core of corpus linguistics, so its frequent recourse to statistical analysis is not surprising.

In this section we simply do not have the space available to provide readers with a working knowledge of the statistics that corpus linguists use, though readers interested in exploring this area should see Oakes (1998), Baayen (2008) or Gries (2009b). Rather, we will highlight some general techniques used in corpus linguistics and, where relevant, certain known problems with the applications of particular statistics in corpus linguistics.

2.6.1 Descriptive statistics

We have noted that most studies in corpus linguistics use basic *descriptive statistics* if nothing else. Descriptive statistics are statistics which do not seek to test for significance. Rather, they simply describe the data in some way. The most basic statistical measure is a *frequency count*, a simple tallying of the number of instances of something that occur in a corpus – for example, there are 1,103 examples of the word *Lancaster* in the written section of the BNC. We may express this as a *percentage* of the whole corpus; the BNC's written section contains 87,903,571 words of running text, meaning that the word *Lancaster* represents 0.013 per cent of the total data in the written section of the corpus. The percentage is just another way of looking at the count 1,103 in context, to try to make sense of it relative to the totality of the written corpus. Sometimes, as is the case here, the percentage may not convey meaningfully the frequency of use of the word, so we might instead produce a *normalised frequency* (or relative frequency), which answers the question 'how often might we assume we will see the word per *x* words of running text?' Normalised frequencies (*nf*) are calculated as follows, using a *base of normalisation*:

$$nf = (\text{number of examples of the word in the whole corpus} \div \text{size of corpus}) \times (\text{base of normalisation})$$

So, if we want to see how often *Lancaster* would be expected to occur, on average, in each million words of the BNC, we set the base of normalisation to 1,000,000 and the calculation would be as follows:

$$nf = (1{,}103 \div 87{,}903{,}571) \times 1{,}000{,}000$$

In this case, $nf = 12.55$. Normalised frequencies based on 'occurrences per thousand words' or, as here, 'occurrences per million words' are the most commonly encountered in the literature; many corpus search tools generate these figures automatically. Note that, in fact, a percentage is simply a type of normalised frequency, where the base of normalisation is 100.

Corpus linguists often compare two or more corpora. Obviously generating normalised frequencies for the corpora being compared is essential when doing this. If we look up *Lancaster* in the BE06 corpus (Baker 2009), we find *Lancaster* occurs only ten times. BE06 is only 1,146,597 words in size, so the raw frequency count does not tell us if the word is more or less common in BE06 compared to the 1,103 instances in the BNC. A normalised frequency does tell us this – the word occurs 12.55 times per million words in the BNC but 8.72 times per million words in BE06. We can calculate a *ratio* to indicate how many times more often the word occurs in the BNC. This is done by dividing the larger number by the smaller – this results in the ratio 1.44, allowing us to say that for every occurrence of the word *Lancaster* in BE06 corpus, the word occurs 1.44 times in the written section of the BNC.

A special type of ratio called the *type–token ratio* is one of the basic corpus statistics. A *token* is any instance of a particular wordform in a text; comparing the number of tokens in the text to the number of types of tokens – where each *type* is a particular, unique wordform – can tell us how large a range of vocabulary is used in the text. We determine the type–token ratio by dividing the number of types in a corpus by the number of tokens. The result is sometimes multiplied by 100 to express the type–token ratio as a percentage. This allows us to measure vocabulary variation between corpora – the closer the result is to 1 (or 100 if it's a percentage), the greater the vocabulary variation; the further the result is from 100, the less the vocabulary variation. Since the size of the corpus affects its type–token ratio, only similar-sized corpora can be compared in this way. For corpora that differ in size, a normalising version of the procedure (*standardised type–token ratio*) is used instead.

You may have noticed something troubling about the statistics we have presented so far. While without doubt they are useful, interpreting them literally leads to absurdity. For example, in the case of the standardised frequency for the word *Lancaster*, it would be foolish to imagine that, if we chopped the BNC into 1-million-word chunks, we would with complete regularity find 12.55 occurrences of the word in each chunk. We clearly would not. The words in a corpus are not, in general, smoothly distributed through it, occurring at regular and precise intervals. In the case of the word *Lancaster*, we would probably expect to find lots of instances bunched together in a small number of texts where the city of

Lancaster is an important topic, and very few instances in the rest of the corpus. Normalised frequencies abstract from, and simplify, the reality of 'what's there' in the corpus. For this reason, it is usually considered good practice to report *both* raw and normalised frequencies when writing up quantitative results from a corpus.

2.6.2 Beyond descriptive statistics

To better understand and characterise the frequency data arising from a corpus, corpus linguists appeal to statistical measures which allow them to shift from simply describing what they see to testing the *significance* of any differences observed. Most things that we want to measure are subject to a certain amount of 'random' fluctuation. We can use *significance tests* to assess how likely it is that a particular result is a coincidence, due simply to chance. Typically, if there is a 95 per cent chance that our result is *not* a coincidence, then we say that the result is significant. A result which is not significant cannot be relied on, although it may be useful as an indication of where to start doing further research (maybe with a bigger sample of data).

The two most common uses of significance tests in corpus linguistics are calculating keywords (or key tags) and calculating collocations. To extract keywords, we need to test for significance every word that occurs in a corpus, comparing its frequency with that of the same word in a reference corpus. When looking for a word's collocations, we test the significance of the co-occurrence frequency of that word and everything that appears near it once or more in the corpus. Both procedures typically involve, then, many thousands of significance tests being carried out. This is all done behind the scenes in those tools that support keyword and collocation extraction. When we wish to apply significance tests to *other* quantitative data extracted from a corpus, however, we cannot normally count on the analysis software to handle the details for us; we must carry out the procedure ourselves.

Different significance tests can be used, depending on what type of data we have, for instance the *chi-square* test, the *t-test* and the *log-likelihood* test (Dunning 1993). The results of these tests allow researchers to assert with a degree of confidence that the results of their analysis either are or are not significant. This provides a distinct advantage to the corpus linguist – statistics, rather than simply describing the data in the corpus, can begin the process of sorting out significant differences from non-significant differences, thereby focusing the work of the linguist in accounting for and explaining the data in front of them. However, the use of significance tests is, predictably, not quite as straightforward as this. Some statistical tests make certain assumptions about the data. For example, the chi-square test presupposes a so-called *normal distribution* of the data. Data has a normal distribution if most of the values cluster relatively tightly around a mean (average) value – a pattern which, when plotted on a graph, gives us the classic 'bell-shaped' curve. This is not true for language data. For example,

word frequencies do not follow a normal distribution. Rather, they typically produce a markedly positively skewed distribution, with a pronounced 'hump' of a few very high-frequency words, followed by a very long tail of lots and lots of low-frequency words. The log-likelihood test (Dunning 1993) is now preferred by some corpus linguists as it makes no assumption of a normal distribution. This is only one example of the known issues with the statistical tests available. Another is that the number of observed examples you have is crucial in some tests, with chi-square becoming unreliable when the number of examples is very small. Again, there are proposed solutions to such problems – including using different statistical tests, notably Fisher's Exact Test (see McEnery *et al*. 2006).

Another complication is that when we do multiple significance tests, we expect some of them to give a false result, just by chance. If we use the standard 95 per cent cut-off point for significance, we expect such a chance result on average one time in twenty. This is potentially a major issue for corpus linguistics since (as noted above) the most common applications for significance testing, namely keywords and collocations, involve thousands of tests at a time. Typically, this problem is addressed by raising the bar for deeming a finding significant. Though 95% is standard in statistics, when keywords especially are calculated, much more stringent cut-off points are used, such as 99.9% or 99.99%. In fact, the default cut-off point for keywords in WordSmith is 99.9999%! Even at such a high level of significance, in most cases plenty of words are still found to be significant keywords.

A general principle emerges from these observations. Much as with descriptive statistics, the *inferential* statistics that allow us to test significance should be used with caution. Taking advice from a statistician, or at the very least consulting books such as Oakes (1998) or Woods *et al*. (1986) to gain an understanding of the limitations, if not the intimate workings, of the test you wish to use, is good practice for any researcher.

Taken together, significance tests form one of the groups of advanced statistics used most often in corpus linguistics. The other main group consists of techniques for *exploring* quantitative data, and investigating structure and relationships within it, rather than testing the significance of a particular result. Many such exploratory procedures exist. Perhaps the most important vis-à-vis corpus linguistics are *factor analysis* and *cluster analysis*. Factor analysis can be used when a large number of different quantitative measures have been made on a particular corpus (for instance, the frequencies of many different grammatical features). The purpose of a factor analysis is to determine which of a large number of quantitative variables are related to each other – a pair of variables are considered related if a change in one always means a change in the other. By determining relationships among variables, factor analysis reduces the number of variables that need to be taken into account (on the principle that two variables that are strongly related to one another are really only one factor of variation). Factor analysis was introduced to the mainstream of corpus linguistics by Biber (1988), who uses it as the basis of the multi-dimensional (MD) method. We will explore the MD approach in detail in Chapter 5. The other commonly used

exploratory technique is cluster analysis, which also investigates the structure of a large group of measurements. For example, imagine you have compiled several different statistics on the usage of a set of different verbs – how often they are used transitively; how often they are past tense; how often the subject is a human being; and so on. It would be quite tricky to tell, just by inspecting your tables of figures, which verbs are similar to one another. Two verbs having very close scores on one measurement doesn't mean they will necessarily be similar on other measurements. Cluster analysis groups the data – the verbs, in this case – according to similarity, taking *all* the information into account. Not only are clusters generated ('clusters' being groups of things, here verbs, that are statistically similar), but the statistical relationships between the clusters are also extracted, showing which are close together and which are relatively further apart. While there are many different mathematical techniques for clustering data, the general principle, of grouping the data according to similarities, is the same in each case. Cluster analysis has been employed for a number of different purposes in corpus linguistics; one notable recent approach is that of Divjak and Gries (2006), who use clustering to produce what they call *behavioural profiles* of words with multiple senses or sets of semantically similar words (see section 7.5.2). The example we gave above, of clustering verbs according to usage features, was effectively a toy example of Divjak and Gries' procedure.

This review of statistics in corpus linguistics is far from complete. However, it does at least serve to show the basic types of statistical analyses that are used in corpus linguistics – descriptive statistics, significance tests and exploratory techniques – and as such will provide a useful, if basic, guide to some of the studies discussed later in this book which draw upon these analyses. In particular, collocation as discussed in Chapter 6 draws heavily on significance tests, making the brief review of it presented above, at the very least, a necessary prerequisite to that discussion.

2.7 Summary

This chapter has reviewed the annotation and exploitation of corpus data. In doing so we have explored a number of objections to the use of corpus data in linguistics. Similarly, we have touched upon objections to the use of annotated corpus data. In both cases our defence is simple – the use of corpora and the use of annotations is not simply justifiable, it represents good scientific practice. As the telescope is indispensable to the astronomer, so the corpus is indispensable to the linguist. As keeping a careful record of analyses is important in any science, so it is in linguistics. However, as important as the use of the corpus in linguistics is, it is worth noting that other methods may also be gainfully combined with or used in place of the corpus – an idea to which we will return in Chapters 8 and 9. The key is to understand the strengths and weaknesses of a range of methods and to deploy them singly or in combination as appropriate. Finally, as has also been

discussed, the tools and statistical procedures that may be used with the corpus are important in determining how useful it can be. They support the exploration of specific research questions with a corpus, and in their absence the usefulness of the corpus itself is either severely constrained or negated.

In the next chapter we will explore two further areas which are important in the construction and use of corpora, namely legal and ethical issues. In discussing legal issues we will focus on building corpora from sources on the web. This will lead to a discussion of some of the advantages and disadvantages of the web as a source of corpus data.

Further reading

Surprisingly little has been written which focuses exclusively on the development of the tools of corpus linguistics. Some books have dealt with the basics of manipulating corpora, such as Barnbrook (1996) and McEnery *et al*. (2006). Other books have focused upon programming techniques that may be used to exploit corpora, notably Mason (2001) and more recently Weisser (2009) and Gries (2009a). However, there are no full-length works which deal comprehensively with markup conventions for corpus linguistics, perhaps understandably: the topic is rather dry and the techniques which are used are often generic techniques for which a literature exists already, as is the case with XML for example. But writing at a shorter length, Burnard (2005) and Wynne (2005) explain some of the benefits associated with the use of XML for corpus formatting, and, in the case of Burnard, provide some examples.

While it is perhaps unsurprising that not much has been written on markup conventions in corpus linguistics, it is less clear why relatively little has been produced looking at the process of corpus annotation. Many papers exist which describe individual corpus annotation schemes (e.g. Brugman *et al*. 2002; Hardie *et al*. 2009) or tagging systems (e.g. for part-of-speech tagging, see DeRose 1988; Brill 1995; Karlsson *et al*. 1995). But the only book which seeks to give the topic a relatively comprehensive treatment is Garside *et al*. (1997). Accordingly this book is strongly recommended as a reference point for the beginnings of a deeper exploration of corpus annotation. Van Halteren (1999) is less broad, addressing only one form of annotation (part-of-speech tagging), but is much more advanced and comprehensive within that purview.

The use of statistical techniques in corpus linguistics is increasingly common, and readers who wish to explore corpus linguistics further should certainly acquire a basic grasp of descriptive statistics and at least seek to understand the principles behind, if not the mechanics of applying, some of the more advanced statistics mentioned in this chapter. Oakes (1998) remains a good resource for those interested in statistics in corpus linguistics, though the book is best used as a reference resource and guide; while it serves as an excellent introduction to many statistics currently used in corpus linguistics, unless the reader is seeking

an extremely detailed understanding of the area, we would advise against a cover-to-cover reading. More recent books on the topic, notably the work of Baayen (2008) and Gries (2009b; see also Gries 2010a), should also be used in the same way by the great majority of readers.

Practical activities

(A2-1) In BNCweb/CQPweb, the query to find 'all words in the corpus ending in *-ness*' is

***ness**

Find out what search pattern you would need to use for the same query in another concordancer you have access to, such as Mark Davies' corpus.byu.edu interface, or WordSmith or AntConc.

(A2-2) Optional or alternative queries can be used in many ways – one use is to take into account the possibility of spelling variation, for example *colour* versus *color* in British and American English. In your concordancer, how would you create a search to retrieve all examples of both *colour* and *color*? (Note: in most concordancers, there is actually more than one way to do this.)

(A2-3) Searching for part-of-speech categories and other tags is often a bit more involved than searching for wordforms or phrases. Get a small amount of tagged data (or tag a raw text file of your own using a system such as the CLAWS web-tagger)[21] and use a third-generation concordancer to search for a word, such as *record* or *convert* in English, that can be a noun or a verb, depending on context. How can you control whether you find nouns, verbs, or both?

Note: depending on what concordancer you are using and what format your tagged data is in (e.g. XML versus *word_TAG* format), you may need to adjust the concordancer's options as well as your search pattern.

Questions for discussion

(Q2-1) Imagine you want to make use of a spoken corpus as part of a piece of research into pragmatics. One common way to go about this is to manually add some pragmatic annotation to the corpus as the first stage of analysis. What pragmatic features might you choose to annotate in a spoken corpus? What kind of classification scheme could you use to label these features? And how could you encode your labels into the corpus? (Hint for those unfamiliar with pragmatics: think about things like speech acts and politeness!)

(Q2-2) Part-of-speech tagging has, usually, around a 3% to 5% error rate. In what kinds of situation might this be a problem when doing research with a tagged corpus? How could you make allowance for these problems?

(Q2-3) Think of as many examples as you can of syntactic structures that have more than one possible analysis in English or another language you know well – either because they are ambiguous, or because different theories of grammar account for them in different ways. In a text with constituency parsing (i.e. phrase-structure brackets), what implications for the placement (and type) of phrase-boundaries would these differences of analysis have? Can you think of any cases where corpus searches or frequency counts might be altered by the decisions made about such structures?

(Two examples to get you started: (a) in *there-is*-type sentences in English, the *there* can be analysed as a subject noun phrase, a clausal adverbial or a unique 'existential' marker; each of these analyses could imply a different layout of parsing brackets! (b) subordinate clauses beginning in *where* can be analysed as adverbial clauses or relative clauses, depending on context, but different analysts draw the boundary in different places.)

3 The web, laws and ethics

3.1 Introduction

In this chapter we will continue our survey of practical issues that may arise when working with corpus data. While the first two chapters discussed the selection, annotation and exploitation of the data in a corpus, in this chapter we will consider two important and related issues: legal considerations in corpus construction; and the equally important, yet less often discussed, ethical issues arising from corpus construction, distribution and use. We will begin with the legal issues, which have become more pressing over time as vast amounts of textual data have become available to collect easily over the World Wide Web. Accordingly, in this chapter we will approach legal issues in corpus construction with specific reference to compiling corpora from the web. In doing so, we will also consider some of the practical issues around web-based corpus construction.

We should note that we write, in this chapter, very much from the standpoint of Western culture. Laws and ethics vary across the planet – but rather than attempt a global survey, we seek here to illustrate the relevant legal and ethical issues from the context in which our own research is undertaken.

3.2 The web and legal issues

The most fundamental issue in corpus construction is whether or not you have the legal right to gather and distribute the data you intend to include in your corpus.

The massive expansion of the World Wide Web in the mid-to-late 1990s presented both opportunities and problems for corpus builders. Before the age of the web, to collect a text in electronic form it was necessary either to get the original file from the publisher, or to rely on re-typing (time-consuming and expensive) or optical character recognition software (error prone). However, the hypertext documents that make up the web are already in electronic form, and frequently in an encoding and format (ASCII text with HTML markup) very similar to the XML format preferred for corpus data. Thus, it has become

extremely straightforward simply to download and save large quantities of text from the web to create a corpus – either manually, or for a larger corpus using an automated program called a web crawler. One such automated program which is specifically designed for linguistics is BootCat (Baroni and Bernadini 2004). This program is 'a suite of Perl programs implementing an iterative procedure to bootstrap specialised corpora and terms from the web, requiring only a small list of "seeds" (terms that are expected to be typical of the domain of interest) as input' (Baroni and Bernardini 2004: 1313). Given the availability of such tools, it is hardly surprising that the study of the 'Web as Corpus' (which we have introduced already in section 1.4.2) has become a highly active subdiscipline of corpus and computational linguistics, with some studies focusing upon genres unique to the web, e.g. online chat rooms (Claridge 2007; Thelwall 2008; King 2009). However, while there are linguists who point out the obvious attractions of the web, conceiving of it as 'a fabulous linguist's playground' (Kilgarriff and Grefenstette 2003: 333), there are others who urge caution, noting that the web 'can in no way be considered a representative sample of language use in general' (Leech 2007: 145). It is unsurprising then that some have concluded that although the web *can* be useful, the 'more sophisticated needs of the working linguist may be better fulfilled by means of traditional corpora' (Lew 2009: 298).

While being able to use the web is a great advantage for corpus construction, there are also problems. One issue is the difficulty of determining the genre of any given document harvested from the web without actually reading it – knowing the genre is necessary for a balanced corpus design. Though BootCat tries to do this, its success is heavily dependent on the user being able to select terms for their search which are strongly associated with the genre in question. An initial evaluation by Baroni and Bernardini (2004: 1315) suggests that one in three of the webpages recovered may not be in the desired genre. However, while the usefulness of the web for linguistics may be debatable, and while the programs, such as BootCat, can be improved, one constant remains – the legal issues are as complex for large-scale web harvesting as for any other type of corpus construction. Copyright laws apply to documents available on the web exactly as they do to print documents. That is, it would be illegal to download a text from the web and then redistribute it as part of a corpus without the permission of the author of the webpage. While this may seem unreasonable, given that the majority of websites are entirely open to public view, many content providers on the web are reliant on fees for advertising that are paid per visitor. So if even one person who might have looked at the original webpage *instead* sees a copy in a corpus, the creator of the original content suffers a financial loss. There can be no objection to someone downloading a single copy of a document on the web onto one computer for their own use (indeed, such copying happens every time a web browser visits a page). But it would clearly breach copyright to redistribute these local copies. This is a serious problem: just as corpus tools need to be widely available if corpus linguistics is to be replicable, the actual corpus data also needs to be made as widely available as possible, for

precisely the same reason. This is, we would argue, an ethical imperative for the researcher.

There are several ways of addressing the copyright issues around data collected from the web. The first is to treat text from the web in the same way as any other text. That is, the corpus builder contacts the copyright holder and requests permission to redistribute the text within a corpus under the terms of some specified licence. This procedure was followed by the builders of pre-web corpora such as the BNC and LOB, but also by the builders of the EMILLE corpora, for which much of the data was gathered from selected websites (see Baker *et al*. 2004). This is feasible if one or a small number of websites are to be sampled. This need not mean a small corpus – many news sites, blogs and internet forums contain many millions of words of archived data. But it would certainly lead to a corpus that represents only a narrow slice of the variety of the web. While this might be a perfect dataset for some sets of linguistic questions, it would be of rather less use for addressing questions about a given language in general. An alternative is to collect data only from sites which explicitly allow the reuse and redistribution of text. A website may declare that its content is public domain, or that it is available under a licence which permits copying and redistributing. This is increasingly common. For example, all pages on the multilingual Wikipedia site are licensed as 'free documentation', meaning that making and distributing copies is permitted (with some conditions). However, restricting a corpus to such sites would again skew its representativeness.

The third approach is to collect data without any regard to seeking permission, and not to distribute it, but instead to make it available to other researchers through a tool that does not allow copyright to be breached. Many web corpora are made available through fourth-generation, web-based concordancers (see section 2.5.4) where only a few words of context around the node word are visible. Indeed most web search engines might be viewed as primitive concordancers of this sort. Since it is impossible to reconstruct the original texts from the tiny snippets in the concordance, which are small enough to count as 'fair use', this 'redistribution' does not constitute a dangerous copyright violation. Thus, the corpus is available to other researchers, since they can run the same searches from the same web-based tool, but the law has not been broken. This is less than entirely ideal, however, because some more sophisticated corpus analyses (especially those involving advanced statistical calculations, e.g. collocation strength across variably defined subcorpora) cannot be done without access to the text of the entire corpus. Davies, in a discussion of such online interfaces, acknowledges this problem, but points out that 'in many cases it is still possible to obtain large amounts of word frequency and n-gram information from these corpora' (Davies 2010: 417) as well as perform advanced lexical and grammatical searches. As fourth-generation concordancers such as Davies' interface continue to be improved and developed, the analyses they permit will be continually extended. But, inevitably, a web-based concordancer will never allow the full range of analyses that a technically savvy researcher could accomplish with a copy of

the corpus on their own computer. Moreover, fair-use law – like all parts of copyright law – may vary between countries (and across legal jurisdictions within countries). So far as we are aware, there has been no definitive establishment of the legal status of fair-use-snippet web concordancers – for instance, no test case has been taken to court that we know of. There are several reasons for this: those who create web concordancers are very careful on legal issues; corpus linguistics is fairly obscure in the grand scheme of things and most text producers probably don't even know if their text ends up in a corpus that is searchable online; and, ultimately, corpus linguists are unlikely to have enough money to be worth suing.

A more innovative solution to the copyright issue is to redistribute not the downloaded data files, but rather a list of the web addresses from which the corpus has been collected (and, if necessary, details of the procedures used to download and process the texts). This does not breach copyright at all – but any researcher with the appropriate software can download those webpages and reconstruct their own personal copy of the corpus from the address list quickly and easily. This ensures that any findings from such a corpus are open to being checked and replicated, at least as long as those pages remain available and unchanged on the web. However, the permanence of webpages is, of course, highly variable; so while this approach has some advantages, it is not a complete solution.

3.3 Ethical issues

While the legal issues involved in corpus construction have been considered widely, less consideration has been paid in the literature to ethical issues in corpus construction. It has been discussed by corpus linguists in forums such as the *Corpora* mailing list,[1] and some authors have directly considered their work in relation to ethical issues, for example Hasund (1998), Sampson (2000) and Rock (2001). However, a number of major works on corpus linguistics, including Sinclair (1991), Kennedy (1998), Biber *et al*. (1998), and McEnery and Wilson (2001), do not treat ethical issues in any depth. This may be because there are existing ethical guidelines for gathering and using data that have been developed by professional linguists and these are used routinely by corpus linguists. For example, the British Association of Applied Linguistics has a well-developed set of ethical guidelines which are clearly relevant to corpus builders.[2] In addition, most universities and other research organisations have their own internal ethical guidelines and procedures that researchers must follow. Nonetheless, a survey of the corpus literature reveals precious few examples of explicit discussion of ethical issues in corpus construction and use. Given that one or two corpora seem not to be wholly in line with ethical best practice, as will be shown, we might reasonably conclude that research ethics is an area that corpus linguistics should consider in more detail. Ethical issues in corpus linguistics can broadly

be divided into four main areas: ethical issues affecting respondents in a spoken corpus, ethical issues affecting corpus builders, ethical issues affecting corpus distributors and ethical issues affecting corpus users.

3.3.1 Ethics and respondents

What type of ethical issues affect respondents? The spoken part of the BNC provides a good example. When this data was collected, many people generously agreed to have their speech recorded in naturalistic settings (Crowdy 1995). The result is a very useful dataset indeed. However, the people carrying those tape recorders around had to give informed consent – they had to understand that whatever they said when that tape recorder was running might well eventually be used in a corpus that would be available to all who cared to use it. They were sacrificing their privacy.

Not only the privacy of what is said in their conversation, but also the privacy of personal information, may be sacrificed by a respondent. In particular, they may provide information about themselves to the corpus compiler which is useful in the generation of demographic metadata but is not to be incorporated into the metadata itself. In this case, keeping that data secure in perpetuity is an important responsibility that a corpus builder must discharge. For example, a speaker may reveal their occupation and workplace to allow their social class to be determined, on the express understanding that such information is not to be passed to a third party. Or they may give a detailed account of all the places they have ever lived, to assist in the classification of their dialect. Preserving the original information is important, so that the social categories assigned to the respondent can be validated when necessary – yet ensuring that access to this data is on a strictly need-to-know basis is just as important.

The ethical imperatives arising from the respondents' sacrifice of privacy were clearly addressed by the builders of the BNC. But an equally pressing ethical issue relates to the people the respondents spoke to on tape – they *also* sacrificed their privacy, and it was therefore necessary for them to give their consent in an equally fully informed fashion. This consent was collected by the respondents themselves; so we see that in this case the process of data collection has actually imposed an ethical obligation *on the respondents*. This raises complex and not entirely soluble problems. Respondents are, of course, not themselves researchers and can hardly be expected to accept the full weight of the ethical obligations that rightly adhere to researchers. In fact, we would argue, when spoken data collection (unavoidably) imposes such obligations on respondents, it only makes the researchers' responsibility of ethical oversight that much more critical. It is the researchers who must ensure and guarantee that good ethical practice is followed by each respondent – and, if there is any doubt about whether consent procedures have been followed fully and correctly by any respondent, the resulting data cannot ethically be included in a corpus.

Less obvious than the privacy of speakers in a corpus, but nonetheless important, is the privacy of the people *talked about* in the corpus. Their privacy is also arguably breached at times. Sampson (2000: section 4.1) summarises the problem well when he notes that in the BNC, two speakers:

> comment that one of their schoolmates, identified by Christian name, behaves like a whore. This person is entitled to anonymity as much as the speakers, and arguably more so: she signed no release form for the corpus compilers. When well-known public figures or institutions are mentioned, the BNC compilers seem to have felt that there was no need to anonymise the references at all. Clearly, if someone announces that he has just bought the latest album by a named pop singer, there is no point in concealing the singer's name. But it depends on what is said. One of the CHRISTINE texts contains a series of quite damaging remarks about the management of a secondary school, named in the BNC file. In another case, speakers comment adversely on the sexual morality of a named American actress. Even American actresses, surely, are entitled to have their honour guarded by corpus linguists.

It would be much more difficult to obtain the consent of people who are not participants but are merely discussed in a given conversation. Accordingly, the privacy of all should be protected by anonymisation – names should be changed, though this may be done in a way which retains linguistically useful features of the name (e.g. a name typical of a female should be substituted with another name typical of a female). In addition to names, however, other personal data should be anonymised – there may be references to home addresses in a corpus, for example. The following example from file F86 of the BNC (utterances 264 and 265) shows how superficial anonymisation may be when sufficient information is left in the text to allow it to be circumvented. In this example, surnames have been deleted to provide anonymisation (as shown by the XML <gap> tags), yet the remaining context is clearly sufficient to allow them to be discovered.

> During nineteen ninety one the Board has been delighted to open new areas of work in Inverness where our first designated place and associated hostel was opened on a most happened– happy day by Sir Russell <gap desc="name" reason="anonymization">.
>
> In Elderslie near Paisley <pause> where Lady <gap desc="name" reason="anonymization"> the wife of last year's Lord High Commissioner opened our fourth senile dementia unit.

Though the example above is relatively benign, we can imagine serious consequences arising from a failure to take privacy seriously in the construction of a corpus. The Speech Act Annotated Corpus (Leech and Weisser 2003) contained credit card details that had to be anonymised, while the Lancaster Corpus of Children's Writing (Smith *et al.* 1998) contained the personal details of young children. The example in Figure 3.1, from file HE7 in the BNC, shows clearly

Speaker	Utterance Number	Utterance
A	275	Well don't you think that it's really rather improper for you to be doing this?
	276	After all people are entitled to some secrecy <event: "running down stairs">aren't they, about their <unclear> <event: "breaking furniture">You don't feel that there's any need at all to give any explanation of your behaviour?
	277	<event: "noise - traffic">You don't think that <unclear> an explanation is due here?
	278	This information after all should have received confidential and does belong to other people, doesn't it?
B	279	What I thinks embarrassing <-|-> is that <unclear> <-|->
A	280	<-|-> And you're just stealing it <-|-> you're just stealing it so that you can make money aren't you?
	281	<event: "noise - traffic"><voice quality: shouting>Do people have a right to have their health records <unclear> confidential do they not? <end of voice quality>
	282	<pause> Have you got nothing to say what so ever?
B	283	'Fraid not, no. <event: "footsteps"> <pause>
A	284	Robert <gap desc="name" reason="anonymization"> is not alone in selling personal information from data banks.

Figure 3.1 *A thief revealed.*

why we may need to consider the privacy of those spoken about in a corpus. It would be problematic if, due to a corpus being published, a conversation like this led to a person being investigated by the police. In a case like this, if the data is not anonymised appropriately it might well be possible, going off what is said in the corpus, to walk around to the house of the person being accused of stealing, knock on their door and show them the transcript. Editing and anonymising such material in a transcript is one matter; ensuring that anonymisation occurs in audio and video recordings is obviously equally important, though much more challenging technically. It is now possible to construct mixed corpora of video, audio and textual material from a range of sources. But when the different media in such a mixed corpus are closely integrated, this can increase yet further the level of detail about specific individuals that is present, making anonymisation of the data yet more problematic; so it is not surprising that such corpora are rarely made publicly available.

The COLT corpus is a good example of a thoughtful approach to the issue of anonymity in corpus building, with due consideration being given to a range of ethical issues over time, as Hasund (1998: 16–17) discusses:

> In the invitation to take part in the research project . . . the following promise was given to the COLT recruits: 'You and the people you have recorded are guaranteed full anonymity.' There were lengthy discussions among the researchers working on the corpus of what was implied by the term 'full anonymity', resulting in an agreement to delete all surnames and addresses in the transcription, but leave all first names unchanged. Considering that the

> recordings were made in a huge city like London, and the recruits were pupils and not public persons connected to specific positions at specific universities of companies, this level of anonymization was considered sufficient for the protection of personal identities.

In spite of this policy, the COLT team eventually decided, after further consideration, that their approach to anonymisation was ethically and legally problematic from the point of view of informed consent. They decided, in a later project focused upon Norwegian teenage speech, to change the statement given to participants to make it clear that their first name would be retained in the published corpus (Hasund 1998: 24–5). This thoughtful, developmental approach to the issues that both the law and ethics present to corpus linguists is clearly what is needed in a field which is relatively new and which is using technologies and methods which are relatively novel. There is, however, a price to be paid for giving due consideration to such ethical issues. The observer effect is undoubtedly amplified by adopting an appropriately ethical stance to the gathering of spontaneous data of this sort. If the respondent knows that they are being recorded, the chance that they will adapt their speech accordingly is undoubtedly elevated.

3.3.2 Ethics and corpus builders

Ethical issues also attach to corpus builders. We will use an example from our own experience. When constructing a parallel corpus of English aligned with a number of South Asian languages as part of the construction of the EMILLE corpus (Baker *et al*. 2004), we were working very much in an opportunistic mode – there was little available data covering all the languages we needed to include in our corpus. We were approached by a religious organisation which wanted to help; they translated many of their magazines and leaflets into the languages in question from English originals. While this represented a golden opportunity to expand our corpus, we felt we had to decline their offer. In part it was because they saw the corpus, which is used in South Asia, as a way of distributing their material and thus gaining converts. We were uncomfortable with the idea of corpus work becoming missionary work. More importantly, when we surveyed the material itself we found the material to be offensive in our view; for instance, one magazine ran an article entitled 'Who it is alright to hate' (*sic*). Our aim was not to construct a corpus of missionary texts. Nor was it to construct a corpus of morally censorious texts. We certainly had no wish to be involved in proselytising. Accordingly, we decided that the data compromised our ethical stance and rejected the offer of the data. This example is at least clear. Our dilemma might have been more acute if we had been working with a sampling frame for which religious texts are required, such as the Brown Corpus sampling frame. In that case, rejecting the material on grounds of offensiveness might arguably have been unethical, since we would have effectively been skewing the balance of the corpus towards particular philosophical or theological perspectives.

This is clearly more a question of research ethics – ethics relating to the conduct of scientific experiments – rather than a matter of the rights of respondents or text owners. It is a useful example to demonstrate the nuanced, multifaceted nature of ethics that corpus linguists should consider.

Returning to the example at hand, we did not feel we had an issue of research ethics in this case as we were aiming only to collect samples of official documents, very broadly defined, and so the issue of skew did not arise. In consequence the offensiveness of the texts became the primary ethical consideration. Any corpus builder may potentially find themselves in a position where they are forced to confront an issue like this, because the underlying problem is embedded in one of the great strengths of the corpus approach: corpora are multifunctional. Once built, they may be used for a wide range of purposes, some of which the builders of the corpus would never have imagined, and quite possibly some which they would never have approved of.

3.3.3 Ethics and corpus distributors

Corpus distributors may face ethical issues of their own. For example, consider that you had built a corpus of texts exploring the language of a radical terrorist group. Would you be content to make your analyses available to the security forces of the country in which the group operated, to help them develop a counter-propaganda campaign? You might or you might not, but there are undoubtedly ethical issues to be considered, especially if you had worked with the radical group to gather the material on the understanding that the texts would not be given to third parties for whatever purpose. Real examples close to this one exist. It might be difficult to conceive of corpora being deployed in support of the US military, but in the USA, much corpus-based work is funded by DARPA, the Defense Advanced Research Agency. On its website is the statement 'DARPA's mission is to maintain the technological superiority of the U.S. military.'[3] Any publically available corpus may be used on DARPA-funded projects, quite legally and legitimately. But the technological superiority of the US military is not a goal that everyone on the planet would support. So here is a situation where a corpus made openly available could end up, without any legal obstacles, being used for purposes that the original corpus builders, and those contributing to the corpus, might never have agreed to. Personally, we do not disapprove of DARPA's use of corpus data, and this is just one example out of many that we could have used. But what it exemplifies is that there are clearly ethical issues to be faced by corpus distributors when they are establishing the rules under which a corpus is redistributed. The offer of religious texts which we discussed above would have posed an ethical problem of a similar sort. Had we built a corpus knowing that a possible outcome of a user reading the material was religious conversion, we would have had to think very seriously about the issue of distribution. Quite apart from our personal ethical considerations, there are countries where seeking converts to a religion is widely regarded as immoral or is, in certain circumstances,

illegal. So at the very least, we would have had to address this issue by warning potential users of the nature and intent of the material when we distributed it.

Another key issue for corpus distributors, whether individuals or organisations, is the responsibility they bear for making sure that the data they hold remains intact and available. As a key goal of corpus linguistics is to aim for replicability of results, data creators have an important duty to discharge in ensuring that the data they produce is made available to analysts in the future. To simply delete data, or to deny it to future researchers, without a very good reason (for example ethical or commercial considerations of the kind that would necessarily take precedence) would be a great breach of trust and certainly represent a dilemma in terms of research ethics.

3.3.4 Ethics and corpus users

Finally, what ethical choices face the user of a corpus? Some may arise from the nature of the analyses. This can be shown very clearly in the case of forensic linguistics (Coulthard and Johnson 2007), an area which has, arguably, grown in power and significance because of the availability of corpus data and the techniques developed by corpus linguists (see McEnery *et al*. 2006: 116–20). In forensic linguistics the choices that the analyst makes are closely tied to a very serious outcome – somebody may be sent to prison unjustly, or a guilty person may walk free. Analysts have to think very carefully in such a situation about the consequences of the analysis they are undertaking, in particular how credible and reliable that analysis is. Normally a corpus analyst may be prepared to accept the possibility of error in an analysis; after all, such errors can be uncovered and corrected by repeating and refining the analysis over a period of time. But in forensic linguistics this approach to the data would be cavalier in the extreme, given the very high, very real stakes involved.

It should also be noted that corpus users have an important ethical duty to make the analyses on which their results were based available to future researchers, again in the interests of replicability. This point intersects with our discussion of corpus annotation (section 2.3). Making analyses clear and available to all is clearly good ethical and scientific practice. It has another dimension, however. Those analyses may be based on algorithms embedded in particular computer programs – if so, maintaining these programs, or at least maintaining a clear record of the detailed procedure by which they operated, may be as important as maintaining the results they produce. While this sounds straightforward in theory, in practice it can be quite difficult. Programs such as part-of-speech taggers are often continuously updated and improved by their developers. A corpus analysed by the CLAWS part-of-speech tagger (Garside *et al*. 1987) in 1989 will have a rather different pattern of errors to a similar corpus tagged by CLAWS six months ago – and unless careful records are kept of which version of a program has been used on each text, this information will be lost. It is entirely possible that, over

time, a researcher might produce (and, critically, publish) different sets of results that are based on the same data, but with different analyses produced by different versions of the same program. So there is a clear need to consider the storage and archiving of these analyses. In this case, ideally the program and its resources should be properly version-managed, and a record of every version archived, in the interests of replicability. The point does not simply apply to automated analyses, however. It applies to all analyses, including manual analyses of the corpus data. If a researcher decides that there are seventy-eight examples of a particular feature in a corpus, then it is incumbent upon that researcher to maintain a record of where and what those examples were. As we argued in Chapter 2, corpus annotation can be of help here as it allows the user to encode the analyses into the actual corpus data.

Another problem that a corpus user faces is that an analyst cannot be sure how their results will be interpreted by others. This is most serious when results are misinterpreted – and the misinterpretation very widely disseminated – by the mass media. The press is notoriously poor at accurately reporting on scientific research (see Goldacre 2008 for many examples). Of course, linguistics is not nearly so often the focus of press interest as the physical, and especially biomedical, sciences. But it would be foolish to imagine that corpus linguistic research is safe from media misinterpretation. For example, research into recent or current language change can often become confused with popular narratives of prescriptivism and language decay. Outside linguistics, it is widely believed that the 'proper' standards are being abandoned and language is becoming sloppier and less correct, a notion which can become interlinked with political anxieties over social decay (see Cameron 1995: 78–115). Press coverage of Leech's research on the changing use of modal verbs (see section 5.3.2), for instance, often viewed the changes that Leech reported through this filter – which led in some cases to condemnation of *Leech himself*![4] In practice, there may not be much a researcher can do to correct mass media misinterpretation of their work. Regardless, we would argue that they have an ethical obligation to *try* – as a part of the more general responsibility of academia to the society that supports it.

3.3.5 Some cases of ethically problematic research

We have given, admittedly, no more than a thumbnail sketch of ethics in corpus linguistics. However, we may now consider how widespread good ethical practices have been in corpus linguistics to date. The literature, while somewhat silent on the question of ethics, generally embodies sound ethical practices. As noted already, corpus users do take privacy seriously when working with respondents. Corpus builders do not routinely produce corpora that are overtly unethical. Corpus distributors often go to great lengths to ensure that the data they distribute is produced to the highest legal and ethical standards – browsing the catalogues of the European Language Resources Association and

the Linguistic Data Consortium provides substantial reassurance to those who wish to consider ethical practices in corpus linguistics. Users of corpora often explore very sensitive issues in a nuanced and responsible manner; work in forensic linguistics is thorough and thoughtful. Linguists such as Baker (2008) explore sensitive issues, such as the representation of paedophiles, thoroughly and fairly. Nonetheless, it is possible to find instances of poor practice. The following examples could reasonably be argued to be in this category.

As noted, the BNC spoken corpus has a somewhat haphazard approach to anonymisation. Names are anonymised. Other features of the transcript which impact upon privacy are not. Worse, the original recordings that make up the BNC are available to the public and have not been anonymised at all – an afternoon spent listening to the recordings at the British Library would allow anyone to start unpicking the anonymisation that does exist in the text. This is not ideal. A response to this problem would be to restrict access to the recordings to ensure that this breach of ethics is limited (as happened with the COLT corpus, Hasund 1998: 16). This was the approach taken to an even more notable breach of ethics: as part of his work prior to the construction of the Survey of English Usage, Randolph Quirk gathered materials through surreptitious recordings (Quirk 1957). This is not necessarily a criticism of Quirk – by the standards of his time he was almost certainly not acting unethically. However, by modern standards this kind of data collection would definitely be judged unethical, and, as enticing as the prospect of using such data may be, it should clearly be shunned by linguists, and no such material should be gathered or used again. Finally, some data from the past is now simply lost – for example, while the work of Phillips (1989) is of great interest, it is quite impossible to replicate on the basis of the information in the published study; neither the algorithm nor the data used are fully presented.

In defence of most of the examples given, corpus linguistics has developed only recently, and the most marked examples of problematic ethical choices are firmly in the past. Some problems are attributable to the unstable nature of computer file storage in the late twentieth century; the ever-changing nature of file storage standards in that period quite understandably, though regrettably, led to a loss of data. With the spoken BNC, the researchers building it in the early 1990s were in *terra incognita* (just as Quirk was in the 1950s). A spoken corpus on the scale of the BNC had never been produced before. In doing something new, they made some mistakes. But the general message is that though corpus linguistics has made mistakes in the past, it has learnt from them, as noted. Nonetheless, ethical considerations remain an important area of development for corpus linguistics. We can only imagine that they will become ever more salient as the data available to the corpus builder grows, and the ability to cross-match it with other digital information sources, such as medical and educational records, emerges. As this capacity grows, however, so will the interaction between corpus linguistics and legislation – for instance, medical and educational records in the

UK and elsewhere are subject to legislative protection. The legal and ethical issues corpus builders have faced in gathering spontaneous speech, for example, may pale by comparison to those that they face when engaging with (meta)data that is subject to specific legislative protection.

3.4 Summary

This chapter concludes the two-chapter overview of what we might call the practical issues faced by corpus linguists. In the following chapters, we will explore different perspectives on corpus linguistics, looking at how they have been used in different traditions of corpus use. The next chapter surveys the work undertaken with corpora in the paradigm of English Corpus Linguistics. It was within that tradition that much of modern corpus linguistics developed, and that many of the issues discussed in this and the previous chapter were first explored.

Further reading

Hundt *et al*. (2007) is an indispensible collection of papers relating to the use of the web in corpus linguistics. The papers cover the philosophy of the 'Web as Corpus' approach but also sound some helpful notes of caution. In addition, the volume contains papers which demonstrate clearly how the web can be used very productively by corpus linguists.

For some criticisms of the Web as Corpus approach, see Leech (2007); for practical problems encountered by analysts when trying to 'clean' extraneous material from webpages for use in a corpus, see Baroni *et al*. (2008) and Hoffmann (2007a, 2007b) (and for a related problem, see the account on Jean Véronis's blog[5] of his heroic efforts to fathom the meaning of the 'counts' given by Google).

Less is available when we turn to the legal and ethical aspects of corpus use. Considering the importance of legal questions when constructing corpora, it is surprising how little has been written on the topic, with most books on corpus linguistics mentioning it in passing, if at all. McEnery *et al*. (2006: 77–9) has a brief section devoted to legal issues which is worth reading. Similarly the issue is addressed briefly by Hundt *et al*. (2007).

If little has been written directly on legal aspects of the use of corpus data, even less has been written on the topic of corpora and ethics, with all of the major works on corpus linguistics being silent on the issue. As noted in this chapter, individual papers have addressed the issue, but no overarching work looking at corpora and ethics has appeared. This is perhaps because, as noted, corpus linguists 'inherit' ethical guidelines and issues from other areas of linguistics, notably applied linguistics. To develop a sense of how corpus linguists do at times confront ethical issues, readers are encouraged to read both Rock (2001) and Hasund (1998).

Practical activity

(A3-1) The next time you find yourself in conversation with a group of friends, imagine that you are secretly recording the conversation. Make a mental note (and, later, a written note) of the ethical issues that might arise if you were intending to transcribe and then publish the data without the participants' knowledge or consent. Repeat this 'experiment' in a number of other contexts. What common issues emerge? Do any ethical issues seem bound to certain types of interaction?

Questions for discussion

(Q3-1) Should a corpus be censored? For example, the BNC has been used to explore swearing in English. This is possible because a choice was made when building the corpus not to censor the data. How defensible is that decision? Are there circumstances in which you would consider censoring corpus data? If so, on what grounds would you do so?

(Q3-2) When we analyse a discourse, we are also contributing to that discourse – especially, but not only, when our results are published. What ethical obligations does this place on an investigator, especially one working on sociologically sensitive discourses, such as representations of a minority group or hate speech?

(Q3-3) Imagine you are building a spoken corpus, and you are collecting data by audio-recording spontaneous conversation. Given that you are following standard ethics procedures to get informed consent from all speakers, what steps could you take, in designing the data collection and transcription procedure, to minimise the observer effect that will inevitably result from following these procedures?

4 English Corpus Linguistics

4.1 Introduction

While corpus linguistics need not restrict itself to any one language, the development of corpus linguistics as outlined in this book was very strongly influenced by work on the English language from the 1960s onwards. In this chapter that work will be overviewed, and its impact upon the development of both corpus linguistics and linguistics in general will be outlined. This chapter consciously takes its title from that of a book edited by Aijmer and Altenberg (1991). It does so because that classic collection of papers summarised the state of the art of that tradition at a key point in its development. Prior to the early 1990s, corpus linguistics was largely the preserve of people working in the tradition of English corpus linguistics. After this point, however, it became more and more a part of mainstream linguistics. It is this tradition of English Corpus Linguistics (abbreviated ECL) that will be reviewed in this chapter, with the progressive development of corpus linguistics being a key topic in Chapters 5 and 6, while its interaction with other kinds of linguistics is the underlying theme of Chapters 7 and 8.

In this chapter, we will argue that ECL has been an important tradition in the history of corpus linguistics. However, while the developments in this chapter were linked to research undertaken on the English language, other languages were developing a tradition of corpus-based study at the same time. Léon (2005) has rightly explored a simplifying Anglo-centric bias in accounts of the development of corpus linguistics, a bias which McEnery and Wilson (2001) sought to avoid. Indeed, the very earliest work that might reasonably be considered to be corpus linguistics, undertaken in the context of humanities computing by Roberto Busa (see section 2.5), was based on a large number of languages other than English. Following on from Busa, other researchers, notably Alphonse Juilland (see, e.g., Juilland and Chang-Rodriguez 1964), built and used corpora for a number of purposes. Juilland developed a wide range of frequency dictionaries based on corpora with a similar balance and representativeness across a number of languages. As with Busa, Juilland's work is notable for its non-English focus. However, although work on non-English corpus linguistics in the latter half of the twentieth century was important, what distinguishes the work in ECL is that ECL was the crucible in which the approach to corpus linguistics

reported in this book and others was formed. The major, systematic contributions of corpus linguistics to the improved description of the lexis and grammar of language were made within ECL. For example, key concepts such as collocation and corpus annotation were developed and refined in this tradition. Importantly, the tools discussed in Chapter 2 were also largely created in the context of ECL. In short, while not all work in corpus linguistics in the latter half of the twentieth century was focused on the English language, it was the network of scholars who did focus on English who largely defined the methods of corpus research as they are currently used in linguistics.

Though the focus of this chapter upon ECL has been justified, a further note of caution must be added. While work on English was influential in the historical development of corpus linguistics as a field, it should not be assumed that, just because the focus of study was English, all the scholars involved were based in Anglophone countries. They were not. While many researchers did work in Australia, New Zealand, the UK or the USA, some researchers came from countries where English is an important second language, for example Hong Kong and India. Other researchers in corpus linguistics at this time came from countries where English is a foreign language. Many worked – and, indeed, continue to work – in western and northern Europe, with Belgium, Germany, Holland, Norway and Sweden making important contributions to the development of corpus linguistics. A number also worked in Eastern Europe; the contribution of researchers in the former East Germany to the development of corpus-based grammars of English is a good example of such work (see Giering *et al*. 1979). In sum, the contribution of scholars from non-English-speaking countries to ECL has been considerable, as this chapter will show.

A final note of caution: it should not be assumed that ECL focuses on one variety of English. It is true that work on British English dominated early work undertaken in this tradition. This was in part because many of the scholars engaged in this work had already been studying British English, so maintaining that focus in corpus-based work was quite understandable. In part, it was also because of the availability of data – the founding of the Survey of English Usage predated the development of the Brown Corpus in the USA, and the corpus-based focus of work on British English was much more sustained. Consequently, the development of British English corpora rapidly outpaced that of American English corpora, so that by the early 1990s it was possible to work on British English looking at hundreds of millions of words of written data using the BNC or the Bank of English, or at spoken corpora such as the London-Lund Corpus (Svartvik 1990). By contrast, at that point the Brown Corpus was still the largest publically available corpus of US English, and no corpora of spoken American English were available. In these circumstances it is easy to see why a focus on British rather than American English may have been amplified by a consideration as mundane as the availability of data. Nonetheless, though British English dominated work in this tradition, the tradition was also sensitive to the need to explore varieties of English. Several important corpora that were aimed at permitting the contrastive

analysis of different global varieties of English were constructed, most notably the International Corpus of English (ICE), as will be discussed later in this chapter. It is also interesting to note that many of the first researchers to construct and use multilingual corpora, such as Karin Aijmer, Bengt Altenberg, Sylviane Granger, Stig Johansson and Tony McEnery, can be identified with the ECL tradition. So while ECL may have been principally concerned with studying British English, this tradition also pioneered a lot of early work on corpus-based contrastive studies, both of varieties of the English language and of different languages.

Work in ECL was typically concentrated in particular research centres, often with key researchers at those institutions being responsible for a formal or informal research group that contributed to corpus-based language studies. Many of these groups were in regular contact with each other through an organisation called ICAME (the International Computer Archive of Modern English),[1] founded in the 1970s. As Leech and Johansson's (2009) account of the formation and early years of ICAME relates, the organisation provided an important framework for a network of scholars to cooperate in the development of ECL as a field. Today, ICAME produces a journal[2] and organises an annual conference where researchers in ECL meet and exchange ideas. The scope of the conference has broadened over the years, and rather than being simply an ECL conference, ICAME now engages with corpus-based studies where English is considered, for example, in contrast with other languages. A sequence of books were produced in loose association with the conferences, in a series entitled Language and Computers.[3] This book series, started in 1988, provides a useful guide for those interested in the development of the field of ECL and, by implication, of corpus linguistics in general. Importantly, ICAME also collected and distributed English corpora. A further contributing factor to the rapid development of ECL, and studies focused on British English in particular, is that not only were such corpora produced more frequently, but they were also among the first to be placed in an archive allowing relatively easy access to the resources by researchers other than their creators.

Linked in to ICAME to greater or lesser degrees were a number of researchers and research centres that interacted to define, in broad outline, what we think of as corpus linguistics today. In the following sections, we will consider the contributions of a number of research centres worldwide who participated in the development of ECL. The review is, of necessity, far from complete – it would be impossible to review every centre and to consider every contribution. What we want to do in the following sections is review what we consider the *major* centres and the *significant* contributions they have made to the evolution of corpus linguistics through their work on ECL: University College London, Lancaster University, the University of Birmingham, the Université Catholique de Louvain, the University of Nottingham and Northern Arizona University. An overview of these centres, taken as a set, will broadly characterise the development of ECL, and add substance to our claim that work in this field provided shape and focus to the development of corpus linguistics. This list of institutions is heavily

biased towards the UK; before we begin our review of these centres, it is useful to consider why this is the case. Geoffrey Leech in a discussion with Núñez Pertejo (2006: 154) explains that:

> It was especially in Great Britain that corpus linguistics flourished, because in the United States, the other most populous English-speaking country, there was a kind of intellectual difficulty about corpus linguistics . . . it went contrary to the mainstream of the generative school. Somehow we were able to develop in a way that the Americans were not, and then corpus work was taken up by the publishers of dictionaries, who certainly found that was a good way to develop their lexicographical publications.

4.2 University College London (UCL)

Beginning with the work of Randolph Quirk, the influence of UCL on the development of English Corpus Linguistics was focused principally in four areas: data collection, the exploration of varieties of English, the development of new grammars of English, and the construction and exploration of treebanks.

The Survey of English Usage (SEU), started in 1959, was the first attempt to provide an ongoing collection of present-day English that would, over time, facilitate the diachronic study of British English. It was a precursor of later corpora such as the British National Corpus and the American National Corpus (Ide and Reppen 2004), as it sought to balance its approach to the English language, recording both written and spoken English and sampling them in a range of genres and contexts. The SEU was very much a groundbreaker in corpus linguistics. Initially the corpus was not stored on a computer at all. It was stored on file cards and only later converted into a computerised form, the spoken part of which is available as the London-Lund Corpus (Svartvik 1990). When computerised, the corpus contained 1 million words of grammatically analysed modern British English. Given that the material in the corpus was collected over a thirty-year time period (1955–1985), the corpus can reasonably be described as an early attempt to provide a resource for the diachronic study of contemporary British English. Indeed, the corpus has been re-edited, and an 800,000-word dataset specifically tailored to facilitate the diachronic study of English, the Diachronic Corpus of Present-Day Spoken English (DCPSE), combines and contrasts spoken material from the SEU and ICE.

The team at UCL, led by Sidney Greenbaum, also took the lead in developing what is still to date the largest corpus for the comparative study of varieties of English, the International Corpus of English (ICE; see Greenbaum 1996). This corpus includes a very wide variety of Englishes from around the world, including Australian, British, Hong Kong, Indian and Irish English.[4] The goal is to compile a series of comparable 1-million-word corpora for varieties of English,

representing both written and spoken forms of the language since 1989. Currently the ICE project aims to include twenty varieties of English, though the corpus will doubtless increase in scope and scale in the future, as it has to date. While it is not geared towards the diachronic analysis of English, unlike the Brown Family of corpora (see section 5.3), ICE remains an unequalled resource for the synchronic comparative study of varieties of English.

From its earliest days, one of the features that made the UCL contribution to corpus linguistics distinctive was its engagement with the parsing of corpus data. The materials in the SEU were manually analysed by grammarians, who introduced simple grammatical analyses into the corpus data. Given the salience of grammatical analyses in the UCL approach to ECL, it is not particularly surprising that one of the major contributions by the UCL team to the development of ECL was in the area of grammar production. Arising from the work on the SEU, the *Grammar of Contemporary English* (Quirk *et al*. 1972) and the *Comprehensive Grammar of the English Language* (Quirk *et al*. 1985) were published. The 1985 grammar was also the first widely distributed modern corpus-informed grammar, making its publication something of a milestone in the development of corpus linguistics. Some earlier grammars had drawn on corpus evidence, notably that of Fries (1940) (see also Dons 2004 for an excellent overview of the descriptive adequacy of English grammars in the sixteenth and seventeenth centuries). However, Quirk *et al*. (1985) set the standard for the grammars that followed. As Mukherjee (2006: 337) notes:

> the *Comprehensive Grammar of the English Language* . . . has been widely acknowledged to be the authority on present-day English grammar, bringing together descriptive principles and methods from various traditions and schools in order to cover grammatical phenomena as comprehensively as possible . . .

That is not to say that the *Comprehensive Grammar* remains unparalleled – as will be discussed later in the context of work at Northern Arizona University, the benchmark for corpus-based grammars has now been substantially raised (see Mukherjee 2006 for a discussion and comparison of corpus-based grammars). The *Comprehensive Grammar* contains features which would not be acceptable in modern corpus-informed grammars. For instance, many of the examples given appear to be, at best, modified corpus examples; statistical data is provided only sporadically and generally focuses on very rare or very common features. Nonetheless, by comparison with what went before, the *Comprehensive Grammar* provided a new model for a corpus-based grammar, a model which in a refined form continues to dominate the field of English reference grammars.

A further notable development at UCL came when a team led by Bas Aarts pioneered the annotation of corpus texts with a phrase-structure grammar based on that used in the *Comprehensive Grammar*. That team then went on to develop complex software tools allowing the resulting treebanks to be browsed effectively,

developing a system called ICECUP (Nelson *et al*. 2002) for that purpose (see also our discussion in Chapter 2).

In sum, the contribution of UCL to the development of ECL was substantial. Not only were corpus collection and annotation techniques pioneered at UCL, but the use to which the corpora were put also led, in the case of the *Comprehensive Grammar* at least, to a comprehensive shift in attitudes towards what a grammar of English should look like.

One final contribution of UCL is difficult to understate. It provided a steady stream of grammarians trained in the corpus approach to linguistics. These grammarians went on to develop ECL and to establish much more firmly the methodological basis of corpus linguistics. Notable corpus linguists who gained experience working on the SEU include Geoffrey Leech and Jan Svartvik, both of whom went on to develop corpus linguistics further, at Lancaster University (UK) and the University of Lund (Sweden) respectively.

4.3 Lancaster University

It is unlikely that corpus linguistics would have developed at Lancaster University if Geoffrey Leech had not transferred from UCL to Lancaster. Leech had worked on the SEU at UCL with Randolph Quirk, and brought with him the experience he had gained at UCL in corpus building and annotation. Leech extended the work of the SEU team at Lancaster in three important ways. Firstly, he began to work with computer scientists, both to explore ways in which corpora could be used to develop computer applications and also to see how computer applications might help with the process of corpus building. Secondly, he initiated a process at Lancaster of extending the range of corpus annotation. Finally, due to his collaboration with computer scientists especially, he found that he was able to build much larger corpora than had been possible at the SEU. Each of these points deserves attention, as together they represent something of a turning point for corpus linguistics – enabling the use of corpora by a wider range of linguists than grammarians and lexicographers, and also, crucially, making the corpus of use to subjects beyond linguistics.

Leech's work with computer scientists was defined largely by the working relationship he developed with Roger Garside in the Department of Computing at Lancaster, though Leech did work with other computer scientists, notably those employed at IBM's T. J. Watson research laboratories, such as Fred Jelinek and John Lafferty. Nonetheless, it was Leech's sustained working relationship with Garside from 1970 onwards, as part of UCREL (the University Centre for Computer Corpus Research on Language), that allowed Leech to gain access over time to two important resources: firstly, specialised tools for the editing and searching of corpus data; and secondly, tools for the automated annotation of corpus data. The specialised tools for corpus editing allowed data to be assembled

much faster than otherwise possible. Tools for the correction of automatic tagging, and others which sped up the introduction of manual tagging, enabled annotated corpus production on a scale not previously seen (see Garside and McEnery 1993: 38). Though the editing tools were primitive by today's standards, at the time they allowed Leech's team of grammarians to work more quickly and effectively than had previously been possible. A group of corpora of around 2.8 million words in size, fully part-of-speech tagged and parsed, were constructed and hand-corrected in a period of approximately four years (the effort equalled approximately fifteen person-years for the team as a whole). Similarly, the tools developed on the search side enabled markup-aware searches which, by the standards of the time, were state of the art. These search tools allowed fresh insights into a range of issues, such as how the frequency of a part of speech associated with a word varies by genre (Garside and McEnery 1993), the use and distribution of different forms of demonstratives (McEnery *et al*. 1997), and the degree to which closure is observable in the lexicon across a range of genres (McEnery and Wilson 2001: 165–86).

Prominent among the annotation systems that emerged from Leech's collaboration with Garside was the first viable automated part-of-speech tagging program. While earlier programs had been created to undertake this task (notably by Greene and Rubin 1971), it was not until Garside, Leech and others worked in 1980–1982 to develop CLAWS (the Constituent Likelihood Automatic Word-tagging System; Garside *et al*. 1987) that a fully automated part-of-speech annotation system was developed that worked well across a range of genres. The availability of a robust automated annotation system, which typically achieved an accuracy score of 95 per cent or higher on texts it had not been trained on, led to a fundamental change in corpus building. Rather than taking time to collect a corpus and then even more time to annotate it with basic word-class information, it was now possible to collect a corpus, annotate it while you had your lunch and then work on the annotated corpus in the afternoon, so to speak. This of course assumes that the error level associated with CLAWS was acceptable for the purpose at hand. If the ∼5% error level was not acceptable then, as noted, it was possible to work with the tools available at Lancaster to rapidly correct a large corpus in a matter of weeks, bringing the accuracy of the annotation to a very high level. With the advent of automated annotation and markup-aware editing tools, the process of corpus building was accelerated immeasurably.

Given the availability of corpus annotation tools, it is not surprising that the Lancaster team expanded the range of annotations undertaken on their data. As reported in Garside *et al*. (1997), by the early 1990s the Lancaster team had developed and applied a very wide range of annotations, covering not only written corpora but also spoken corpora (Knowles 1993). Table 4.1 summarises the range of annotations developed at Lancaster.

As well as developing annotations, the Lancaster team also had to develop markup schemes. These were very much 'in-house' schemes to begin with but shifted, as standards became available, to such markup languages as SGML and

Table 4.1 *Corpus annotation research at Lancaster University in the 1980s and 1990s*

Annotation	Reference to a paper on the annotation process	Example of corpus annotated
Part-of-speech	Garside *et al*. (1987)	Numerous, including LOB and the BNC
Prosodic	Knowles (1993)	Spoken English Corpus
Parsing	Sampson (1987)	Lancaster-Leeds Treebank
Semantic	Wilson and Rayson (1993)	The Market Research Corpus
Anaphoric reference	Fligelstone (1992)	AP Newswire corpus
Literary stylistic	Short *et al*. (1996), Semino and Short (2004)	Speech, Writing and Thought Presentation Corpus
Pragmatic	Archer and Culpeper (2003)	A sub-part of the Corpus of English Dialogues 1560–1760

XML. As well as working with new markup schemes, the Lancaster team also developed markup standards for some forms of corpus annotation. This initiative sought to codify experience gained through work on the English language and, taking developments from other centres and work on other languages into account, develop a set of standards for the morphosyntactic and syntactic annotation of corpus data (see respectively Leech and Wilson 1994, 1999 and Leech *et al*. 1995). One final area of innovation in corpus annotation is the work of the Lancaster team in exploring the distinction between the accuracy and consistency of annotation (see section 2.3.2). Baker (1997) undertook an experiment using the trained grammarians in the Lancaster team to see how consistent and accurate the annotations introduced by the analysts were. This was partly in response to a criticism of corpus annotation made by Sinclair (1992), which suggested that while accurate, human annotations might be inconsistent; and partly in response to an awareness in the team that some inconsistency was present in their analyses. Baker's experiment showed that the degree of inconsistency was in fact negligible, while the accuracy was high. This refuted Sinclair's claims, at least for part-of-speech tags introduced by well-trained teams. Work by Marcus *et al*. (1993) and Voutilainen and Järvinen (1995) produced similar results. Such experiments proved influential and further studies of so-called *inter-rater consistency* followed. Indeed, such studies are now a common part of the work of corpus annotators (see, e.g., Peacock 2006).

The final ways in which Lancaster innovated were to work on bigger corpora, and to extend the range of languages for which they were constructed. From 1970, when UCREL was established, until the BNC was built in 1991–1995, Lancaster developed around 106 million words of corpus data. All this material

was part-of-speech tagged, and around 4 million words were parsed. Such a rate of annotated corpus production was unthinkable when corpus-building work began at Lancaster in 1970. While changes in the publishing industry made machine-readable text ever more easily available in that period, it was the work on automated annotation systems and the development of tools to speed manual annotation that allowed this volume of annotated text to be produced. Starting in the early 1990s, the team at Lancaster produced corpora in an ever-wider range of languages, using the techniques and experience developed working on English to produce corpora of Chinese (the Lancaster Corpus of Mandarin Chinese);[5] French, English and Spanish (the CRATER corpus);[6] and fifteen South Asian languages (the EMILLE corpora).[7] This work in turn led to the development of further annotation tools, such as a part-of-speech tagger for Urdu (Hardie 2004, 2005).

While Lancaster produced probably the widest variety of corpus annotations in the period up to the mid-1990s, it was far from the only centre engaged in this enterprise. There were other centres actively treebanking English corpus data at this time. Of particular note in this respect is the University of Nijmegen, where Jan Aarts worked with colleagues such as Nelleke Oostdijk, Pieter de Haan and Hans van Halteren to develop a parsed corpus. Like Lancaster, that group also produced bespoke corpus editing tools, in the form of TOSCA (Tools for Syntactic Corpus Analysis; see van den Heuvel 1988), which allowed them to treebank substantial volumes of data. Similarly, Geoffrey Sampson continued to develop treebanked corpora after he left Lancaster University, notably the SUSANNE corpus (Sampson 1995). Important work on treebanking also took place at the University of Pennsylvania; see section 4.7.[8]

Yet while the production of very large-scale corpora, annotated on a number of levels, remains a distinctive contribution of Lancaster to ECL, the corpora developed at Lancaster were not the largest generated by any centre of ECL. Birmingham, with the Bank of English, produced by far the largest corpus.

4.4 University of Birmingham

While the work at UCL and Lancaster was influential, both in defining corpus linguistics and in nurturing scholars who developed and advanced the field, a quite distinct school of corpus linguistics was developed at the University of Birmingham by John Sinclair. This school, as noted elsewhere in this book, had a distinctive approach to the construction of corpora (see section 1.4.1); questioned the process and utility of corpus annotation (see sections 1.5 and 6.6.3); developed the notion of the collocation as a mainstay of analysis in corpus linguistics, especially in the analysis of meaning (see Chapter 6); and took a corpus-driven approach to the analysis of English grammar, in which words and grammar were ineluctably bound in what has been called *lexicogrammar*

(Halliday 1985, Sinclair 1991 – though Sinclair preferred the term *lexical grammar*). As a distinctive and wide-ranging contribution to corpus linguistics, the work of the Birmingham school will be reviewed in depth later in this book (in Chapter 6). In this section we will simply sketch out the development of the approach associated with Birmingham; review briefly that approach's impact on lexicography in particular; note how that approach produced a number of noted corpus linguists; and, finally, briefly overview some theoretical developments arising from the work of Birmingham-trained corpus linguists, notably Hoey (2005).

Corpus linguistics at Birmingham originated in the 1960s and 1970s (see Sinclair *et al*. 1970, for instance). But it can be said to have come of age in 1980. In that year, Sinclair started a partnership with the publishing company Collins to set up COBUILD (the Collins-Birmingham University International Lexical Database) as a research facility. COBUILD provided data, ideas and analyses for Collins, to help them develop a new corpus-based dictionary (the Collins COBUILD dictionary, 1987). The success of the COBUILD enterprise led, in 1991, to Birmingham and Collins beginning the development of a large monitor corpus, the Bank of English, which continues to grow to this day. To get a sense of the advances that corpus lexicography made possible, consider the following example from Hanks (2009: 216), relating to insights gained by use of the COBUILD corpora in the 1980s into *-ly* adverbs:

> It was widely assumed by pre-corpus lexicographers that all (or almost all) *-ly* adverbs were adverbs of manner modifying the sense of a verb, an adjective, or another adverb and that the meaning of the adverb was always (or almost always) systematically derivable from the root adjective. There is some truth in this, of course: *walking slowly* is walking in a slow manner. But some *-ly* adverbs in English have special functions or constraints, which were not always well reported in the pre-corpus dictionaries. The *Oxford Advanced Learner's Dictionary* (OALD1–4) says nothing about the use of words like *broadly*, *sadly*, *unfortunately*, *luckily* and *hopefully* as sentence adverbs – linguistic devices that enable speakers and writers to express an opinion about the semantic content of what they are saying. Those pre-corpus dictionaries which did notice sentence adverbs did not succeed in noticing all of them systematically.

This is, of course, but one isolated example of a general trend, as Hanks (2009) outlines in detail. Corpus data forced lexicographers to shift away from a focus upon regularities which were easily attested by a few examples towards what Hanks (2009: 216) describes as 'the idiosyncratic conventions that are associated with each word'. As a result of this shift, the descriptive adequacy of the dictionaries produced improved markedly, revealing pre-corpus lexicography to be 'little more than a series of stabs in the dark, often driven by historical rather than synchronic motives' (Hanks 2009: 230). In playing a leading role in bringing about this change, Sinclair undoubtedly made a lasting contribution to lexicography.

As the main scholar associated with the development of corpus linguistics at Birmingham, Sinclair published a number of works outlining his approach to the subject (many of them collected as Sinclair 2004). His work emerged from earlier ideas developed by J. R. Firth (1957). Sinclair also worked as part of a network of scholars which was relatively distinct from the network centred around ICAME, including researchers such as Michael Halliday, whose approach to data shaped and influenced Sinclair's own views (Sinclair 2004: vii); and Michael Stubbs, who developed an approach to the study of discourse in which collocation plays a central role (see Stubbs 1996). In addition, Sinclair collaborated with a number of linguists at Birmingham who, while working in the tradition that Sinclair championed, developed that approach to corpus linguistics further. For example, Susan Hunston has been central in advancing the lexicogrammatical approach to the analysis of language, developing Pattern Grammar, a model where language is built up of a series of linked sequences of fuzzy structures, within which collocation provides both structural coherence and meaning (Hunston and Francis 1999; see also section 6.5.2). Wolfgang Teubert (2004, 2005) has attempted to root the Birmingham approach in older traditions of philology, while Antoinette Renouf (2007) has explored the use of the Birmingham approach in the Web as Corpus paradigm.

Importantly, the linguists who worked with Sinclair helped to spread, and to a degree popularise, the Birmingham approach to corpus linguistics – though Sinclair, a prolific author, was a great communicator of his own ideas. Of particular note was the move of some key researchers from Birmingham to the University of Liverpool in the 1990s. This created for a time a new and vital centre for the study of corpus linguistics along the lines developed at Birmingham. Among the work at Liverpool, of particular note was the development of WordSmith Tools by Scott (1996).[9] WordSmith Tools was one of the earliest third-generation corpus concordancing tools to become generally available (see section 2.5.3). Requiring little expertise and a simple PC to run, the package proved revolutionary when it was published. Importantly, it allowed users to explore collocation in corpus texts with relative ease, whereas previously collocation tools were typically available only at specific sites as part of a first-generation program (see section 2.5.1). It was also at Liverpool that Michael Hoey developed his theory of Lexical Priming. This was the first comprehensive attempt to develop a theory of language on the basis of the Birmingham approach, with Hoey making specific claims about the link between frequency, collocation and language in the mind (see section 6.5.3).

4.5 Université Catholique de Louvain

Sylviane Granger of the Université Catholique de Louvain in Belgium made a notable contribution to the development of ECL when her institution established the International Corpus of Learner English (ICLE) in 1990

(see Granger 1993a, 1993b, 1994). Granger leads an international consortium who are committed to making comparable corpora composed of the English writings of L2 learners of English with a specific L1 background.[10] In order to make the corpora comparable, the students generally produce a series of essays of roughly similar length on similar topics. These topics require the students either to write argumentative essays in response to questions such as 'Is fox hunting justifiable?' or to produce essays as part of a literature exam. To date ICLE contains over 4.5 million words arranged in sixteen subcorpora, each of which contains the writing of students from a distinct L1 background (Granger *et al.* 2009). As well as collecting the text, the corpus also captures some background information on the contributors to the corpus, recording information such as their sex, educational level and their educational experience.

Later work by De Cock (1998) has expanded the study of learner English from written to spoken English with the Louvain International Database of Spoken English Interlanguage (LINDSEI). This corpus is more limited than ICLE, as the first release of the corpus covers only fifty speakers with a French L1 background, and is only 100,000 words in size, though a number of teams are working internationally to expand the corpus.[11] The production of LINDSEI was a pioneering attempt to facilitate the study of learner speech, which has been further enabled by the production of a native-speaker corpus, the Louvain Corpus of Native English Conversation (LOCNEC), which has been used as a control corpus for exploring LINDSEI (see, e.g., Aijmer 2009).

In addition to the production of learner corpora produced by learners, Granger's team also had to engage with the issue of how to compare the material gathered to L1 English material. In response to this, a further corpus was collected, LOCNESS (Louvain Corpus of Native English Essays). This consists of argumentative essays by British and American writers and is some 324,000 words in size. The corpus is divided into three sub-parts – British L1 English writers in pre-university study (17–18 years old), British L1 English writers at university, and American L1 English writers at university.

There is little doubt that Granger's work on learner corpora has stimulated a whole new field of ECL. While work based on what might very broadly be termed learner corpus data was undertaken prior to work on the ICLE project, for example by Dulay and Burt (1973) and Krashen *et al.* (1978), the scope and systematic nature of Granger's work marks it out as the first large-scale engagement with a corpus-based approach to the language of learners. It has generated a great deal of academic output (see Granger *et al.* 2002 and Gilquin *et al.* 2008, for example)[12] and has proven highly influential in the world of English language teaching (ELT) publishing, as well as in studies of second language acquisition (Ellis 2008). The publishers Longman and Cambridge University Press have both developed 'in-house' learner corpora that they now regularly use to inform their publications aimed at the ELT market. For example, the Longman Learner Corpus is 10 million words in size and, like ICLE, is composed of a series of subcorpora of data from students with specific L1 backgrounds producing L2 English writing. Longman uses the corpus to create language learning materials

which are designed to counteract the errors made by students with specific L1 backgrounds:

> The Longman Learners' Corpus offers so much invaluable information about the mistakes students make and what they already know, that it is the perfect resource for lexicographers and material writers who want to produce dictionaries and coursebooks that address students' specific needs. The Longman Learners' Corpus was used to write the Usage Notes in the Longman Active Study Dictionary. Evidence from the corpus showed that there was 100% error in the meaning and the use of the word ***cloth:*** **My cloths and shoes were wet, We have very good cloth stores** etc. Having pinpointed such a problem, our lexicographers were then able to write a corresponding Usage Note . . .[13]

The Cambridge Corpus is some 30 million words in size and is continually growing. It is of particular interest in that annotated within it are the locations and types of the errors made by the learners when writing. The advantages of the inclusion of such annotations are obvious:

> We can see which errors are typical of different learner levels or of particular language groups because all the scripts have information about the first language and English level of the writer. This means that when we produce a book designed for a particular level, e.g. Upper Intermediate, we can look at all the scripts written by Upper Intermediate learners and very easily see exactly what mistakes they make. In this way we can make sure the book contains appropriate help for an Upper Intermediate student.[14]

Error tagging was another development in learner corpus research strongly advocated by Granger (1999, 2003). Yet while Granger has been clear on the potential benefits of error tagging, noting that 'once the corpus is error-tagged, the return on investment is huge' (2003: 10), she has also sounded a useful note of caution about the process of error tagging. Having noted that it is one of the more subjective forms of corpus annotation, Granger (2003: 11) adds:

> It is also important to bear in mind that error tagging, in spite of its numerous advantages, is only concerned with learner misuse. It fails to uncover other aspects of interlanguage such as the under- and overuse of words and phrases, which together with downright errors contribute to the nonnativeness of learner productions.

In short, while useful, error tagging is not a panacea for the problem of nonnativeness in L2 language use. It is one type of information among many that a learner corpus might provide which can be of assistance to the learner or teacher of English.

The expansion of ECL to include learner English was very much pioneered by Granger; much of the work of commercial publishers such as Longman and Cambridge University Press using their own corpora is influenced, either directly or indirectly, by her work, as is the research of other scholars using corpora outside of the ICLE project, such as Tono (2009). However, the learner corpus

approach to using corpora in language teaching is but one of many ways in which corpora have impacted upon research in ELT. Corpora have penetrated many aspects of ELT and have contributed to more specialised areas such as English for Academic Purposes (e.g. Alsop and Nesi 2009) and English for Specific Purposes (e.g. Mohamad-Ali 2007). Corpus-based research in ELT in particular, and language teaching in general, has focused upon issues such as syllabus design (Mindt 1996; Shortall 2007), language testing (Alderson 1996; Taylor and Barker 2008), classroom teaching practice (Amador-Moreno *et al*. 2006), reference and classroom material production, and student-led learning through the so-called *data-driven learning* approach (Johns 1994, 1997; Boulton 2009). There is also a regular conference series, Teaching and Language Corpora (TALC) whose remit is the use of corpora in language teaching. Possibly one of the most significant contributions of corpus linguistics to ELT has been corpora developed with the goal of producing advanced learners' dictionaries, such as the COBUILD dictionary and the *Macmillan English Dictionary for Advanced Learners*. The corpora developed by publishers for these dictionaries have also been employed for other purposes, notably ELT materials such as the *Longman Language Activator* series.

So while Granger's contribution is undoubtedly important, the full impact of corpora upon language learning is now greater than Granger's body of work and encompasses a great mass of diverse research in the general area of ELT, as reported in publications such as Wichmann *et al*. (1997), Burnard and McEnery (2000), Kettemann and Mark (2002) and Reppen (2009).

4.6 University of Nottingham

The work of applied linguists at the University of Nottingham has been influential in English Corpus Linguistics. While Nottingham continues to innovate, establishing itself as a centre for the creation and study of multi-modal corpora (Knight *et al*. 2009), it is for its contribution to the study of spoken English in particular that Nottingham is best known within ECL. Nottingham's work is critical because it explains, in part, how the major grammars of ECL came to focus on both spoken and written English, rather than the written form of the language remaining the (near-)sole object of study for grammarians.

The grammars produced in the 1970s and 1980s by UCL all had the written language as their principal focus, in spite of the fact that a great deal of effort was invested in the production of spoken corpus material by the Survey of English Usage. The grammars of this period were very much rooted in the attitude to speech that casts it as a debased form of language, mired in hesitations, slips of the tongue and interruptions. Also, it should be noted that until relatively recently, there were no large corpora of spoken English available on which to base a study of grammar in spoken English. Some pioneering early work, notably that of

Fries (1940), had used small collections of transcribed speech in order to explore grammar in spoken English, but until the creation of London-Lund Corpus, the spoken section of the British National Corpus and the Santa Barbara Corpus of Spoken American English,[15] there were no substantial corpora of spontaneous speech.

When such corpora were created, a spectrum of opinion regarding grammar in speech developed quite rapidly. One extreme may be characterised as the orthodox position – that grammar in speech is present only in some bastardised form, subject to interference from a host of irrelevant performance features. The advent of spoken corpora allowed an opposite extreme to develop – the view not only that speech is grammatical, but also that it has a grammar of its own which is quite distinct from writing. This latter position is most closely associated with Brazil (1995). Brazil argues for a linear grammar of speech, a grammar which is not sentence-oriented and which does not have 'recourse to any notion of constituency of the hierarchically organized kind' (Brazil 1995: 4) – to put it simplistically, this kind of grammar involves no tree-style parsing. Brazil worked as part of the corpus research group at the University of Birmingham and was heavily influenced by John Sinclair.

The contribution to this debate by linguists at Nottingham, notably Ron Carter and Mike McCarthy, was to begin to move it to the middle of the spectrum. In their work, Carter and McCarthy (1997) used the 5-million-word CANCODE corpus of spoken English developed for Cambridge University Press[16] to explore the nature of grammar in speech. While initially seeming to adopt a position similar to that of Brazil (Carter and McCarthy 1995), they later took an approach that did not call for a distinct grammar of speech. Instead, their approach focused on those features of speech which appear most at odds with grammars taking the written language as their starting point. McCarthy and Carter (2001: 52–3) usefully explore some of the distinctive features of speech using the example from CANCODE given in Figure 4.1.

In Figure 4.1, the bold text highlights a series of issues which grammars dealing with speech would have to engage with (McCarthy and Carter 2001: 53):

> (a) Indeterminate structures (is the second *Take that off* an ellipted form of *I'll just take that off*? Is it an imperative? Is *All looks great* well formed? What is the status of *And they're like*?).
>
> (b) Phrasal utterances, communicatively complete in themselves, but not sentences (*Oh that. For a car. Any problem.*)
>
> (c) Aborted or incomplete structures (*It was a bit erm . . . A bit.*)
>
> (d) 'Subordinate' clauses not obviously connected to any particular main clause (*As soon as they hear insurance claim*)
>
> (e) Interrupted structures with other speaker contributions intervening (*Anything to do with . . . coach work is er . . . fatal isn't it*)
>
> (f) Words whose grammatical class is unclear (*Yow. Now.*)

[Speakers are sitting at the dinner table talking about a car accident that happened to the father of one of the speakers]

<Speaker 1> I'll just take that off. **Take that off.**
<Speaker 2> **All looks great.**
<Speaker 3> [laughs]
<Speaker 2> Mm.
<Speaker 3> Mm.
<Speaker 2> I think your dad was amazed wasn't he at the damage.
<Speaker 4> Mm.
<Speaker 2> It's not so much the parts. It's the labour charges for
<Speaker 4> ~ **Oh that. For a car.**
<Speaker 2> Have you got hold of it?
<Speaker 1> Yeah.
<Speaker 2> **It was a bit erm.**
<Speaker 1> Mm.
<Speaker 3> Mm.
<Speaker 2> **A bit.**
<Speaker 3> That's right.
<Speaker 2> I mean they said they'd have to take his car in for two days. And he says All it is is s= straightening a panel. **And they're like,** Oh no. It's all new panel. You can't do this.
<Speaker 3> **Any erm problem.**
<Speaker 2> **As soon as they hear insurance claim.** Oh. Let's get it right.
<Speaker 3> Yeah. Yeah. **Anything to do with+**
<Speaker 1> **Yow.**
<Speaker 3> **+coach work is er+**
<Speaker 1> Right.
<Speaker 3> **+fatal isn't it.**
<Speaker 1> **Now.**

Figure 4.1 *Spoken transcript from the CANCODE Corpus, from McCarthy and Carter (2001: 52–3).*

By shifting the focus from arguments in favour of the uniqueness of spoken grammar towards the distinctive features of spoken grammar, McCarthy and Carter developed a useful characterisation of spoken English, given full expression in their recent *Cambridge Grammar of English* (Carter and McCarthy 2006). McCarthy and Carter have drawn particular attention to how the grammar of speech can vary by context and can be influenced by the relationship that exists between the speaker and hearer (McCarthy 1998). This observation informed much of the design and development of the CANCODE corpus (Carter 2004) and other corpora developed at Nottingham, notably the Nottingham Multi-Modal Corpus (Carter and Adolphs 2008; Knight *et al*. 2009). This latter corpus has grown directly out of Nottingham's focus on speech, which has led the Nottingham team to look at how speech, gesture and prosody combine to create meaning, following on in particular from the work reported in Schmitt (2004).

The position developed at Nottingham in turn links clearly to the development of later grammars produced in the UCL tradition: if the work of linguists like Brazil, and more importantly McCarthy and Carter, can be viewed as having a distinct influence, it has been on grammars of English. While large grammars of English in the 1970s and 1980s rarely engaged with spoken language, grammars produced since then have done so routinely, and moreover have done so in a way that closely resembles what McCarthy and Carter called for – an acknowledgement of the differences between grammar in speech and grammar in writing. The *Longman Grammar of Spoken and Written English* (LGSWE; Biber *et al.* 1999) is a perfect case in point. This grammar might reasonably be viewed as part of the UCL tradition discussed earlier in this chapter, yet it engages fully with the differences between speech and writing. That said, the LGSWE moves more firmly to the middle ground than Carter and McCarthy did, by arguing that these grammatical differences are largely a matter of degree rather than absolute distinctions. This point is developed by Leech (1998: 11), who explains how he decided to approach the question when he worked on the chapter in the LGWSE dealing with spoken English:

> The first task I set myself was to read through the drafts of all of the preceding chapters, noting grammatical phenomena which were strongly biased in frequency towards the spoken medium. The result of this was a rich profile of conversational grammar as it distinguishes itself from written grammar in all its variety. The profile included some features, such as disjunctive prefaces and tags . . . which had found their way into the 'mainstream' presentation of Chapters 2–4. But in frequency terms, I noted a scale of conversational features going from those which are well represented also in the written medium to those which are virtually absent from it – such as dysfluency phenomena, which in written language are restricted to writing modelled on speech, as in fictional dialogue.
>
> There were on the other hand features which appeared not to find a place in conversation – such as *for* as a conjunction although even here there was the occasional exception. It never seemed realistic, on reflection, to argue that certain features would *never* occur in speech, or would *never* occur in writing, because even if they were not detectable in several million words of conversation or written language (as the case might be), they could crop up if more data were considered. I therefore found myself adopting the 'same grammar' point of view, seeing both speech and writing as making use of the same overall grammatical repertoire, but allowing always for cases where the feature in question might be overwhelmingly commoner in one than the other.

Table 4.2 gives examples from Leech (1998) of features which are either distinctive of speech (what Leech calls *high differential frequency* features), almost exclusively related to speech, or rare in speech and much more common in writing (what Leech calls *low differential frequency* features).

Table 4.2 *Features whose frequency differentiates speech and writing (all examples taken from Leech (1998: 11–13))*

Type of feature	Feature	Example
Distinctive in speech	Front ellipsis deleting subject	*Doesn't matter*
Near exclusive use in speech	Attention signals	*Hey*
	Familiarising vocatives	*honey, mum, guys*
	Omission of auxiliary	*what you doing?*
	Vernacular syntax	*My legs was hurting*
Rare in speech	Dependent genitive	*I met Geoff's student*

Leech (1998: 11–13) explains the differences between speech and writing such as those noted in Table 4.2 according to one of seven features of speech:

1. The presence and use of shared context (e.g. front ellipsis);
2. The avoidance by conversation of the elaboration of specification of meaning (e.g. relative absence of dependent genitive);
3. The interactive nature of conversation (e.g. attention signals);
4. The expressive nature of conversation that allows the communication of personal politeness, emotion and attitude (e.g. familiarising vocatives);
5. The real time nature of conversation (e.g. omission of auxiliary);
6. The restrictive and repetitive repertoire of conversation (e.g. manifested by a lower type-token ratio than writing);
7. The vernacular nature of speech.

In sum, the work of Nottingham in sharpening the focus of corpus linguistics on grammar in speech was an important contribution to the shift from a polarised debate about whether or not speech had a distinct grammar, towards the view now predominant in most schools of English descriptive grammar, where the grammatical system is both flexible and dynamic – where, as Leech (1998: 13) argues, 'English grammar is common to both written and spoken language – but its shape can be moulded to the constraints and freedoms of each.' The work of Brazil, Carter and McCarthy, and Leech is linked very closely to the central concerns of another researcher, Douglas Biber, whose work has also focused on the differences between speech and writing.

4.7 Northern Arizona University and the USA

As noted at the beginning of this overview, corpus linguistics in the 1970s, 1980s and to a lesser extent the 1990s was a strongly European venture, in

spite of the spur provided by the development of the Brown Corpus in the USA by Francis and Kučera (1964). The most notable exception to this trend is observable in the work undertaken initially at the University of Southern California (USC) by Douglas Biber. Biber began his work at USC in 1980, working with Ed Finegan. While Biber began researching phonology, he developed an interest in differences between speech and writing which he started to explore by taking manual counts of features. At the suggestion of Finegan, Biber started using the Brown Corpus[17] and rapidly developed an approach to the characterisation of corpus texts based on the statistical analysis of a range of linguistic features along a series of so-called *dimensions* of variation. We will not explore the technical details of this work any further for now, as Biber's approach is reviewed in depth in the next chapter. The very prominence of Biber's technique in our review of the field is telling of the influence that it has had. Biber's approach to the analysis of corpus data has influenced many researchers and has led to numerous publications using, critiquing and extending his technique, as will be discussed in Chapter 5.

Biber continued this work on the multi-dimensional approach to the analysis of corpus materials when he left USC to work at Northern Arizona University. At Arizona, Biber worked to revivify corpus linguistics in the United States, constructing corpora such as ARCHER (A Representative Corpus of Historical English Registers, developed in collaboration with Finegan)[18] and conducting corpus-based research on a range of languages in addition to English such as Tuvaluan, Korean, Somali and Spanish (Biber 1995a; Biber *et al*. 2006). Biber also took a leading role in establishing the American Association of Applied Corpus Linguistics (AAACL). Like ICAME, this organisation provides networking opportunities for corpus linguists in North America and beyond, in particular through its regular conference series. He also worked with other researchers in ECL, linking the nascent field of corpus linguistics in the USA with that in Europe. The most obvious product of such networking is his work with Stig Johansson, Geoffrey Leech and others on the LGWSE (Biber *et al*. 1999). In doing so he expressly brought American English back into focus in ECL. The LGSWE, as well as innovating in grammar production in a number of ways discussed in the previous section, also innovated in developing a corpus-based comparative approach to the grammar of British and American English.

Through Biber's work, a bridge was formed which reconnected corpus linguistics in North America with the ECL approach to corpus linguistics that had developed in Europe. Another, quite distinct, link to work in North America was formed by computational linguists (see section 9.3), who were interested in corpora primarily as a source of data on which to train and develop language processing systems (e.g. Brown *et al*. 1988; DeRose 1988; Gale and Church 1993). Of particular importance to the rise of corpus-based computational linguistics in the USA has been the foundation of the Linguistic Data Consortium (LDC), a major centre for corpus archiving and production at the University of

Pennsylvania (UPenn).[19] This began with the work of Mitch Marcus and his team, who developed treebanks (see Marcus *et al*. 1993) and annotated corpora in the tradition of those developed by the major European centres (as surveyed in this chapter), which are widely used in computational linguistics. Since then, UPenn has gone on, through the LDC, to become a highly significant centre for corpus production and distribution not simply in the USA, but in the world. While there is little doubt that US pioneers in computational linguistics did much to rehabilitate the corpus as a source of data in the USA, it was Biber who worked on developing distinct corpus-oriented approaches to the study of language in the USA. As such, his contribution in setting up a connection through which ideas crossed in both directions between ECL researchers in Europe and the USA is as substantial as his intellectual contribution.

Since this pioneering work, other notable centres of ECL have developed in the USA, most notably at the University of California Santa Barbara, the University of Michigan and Brigham Young University. The work undertaken at Santa Barbara includes the construction of the Santa Barbara Corpus of Spoken American English, as well as a number of methodological developments, in particular related to developing and testing statistical techniques for corpus analysis (e.g. Gries 2006c, 2008). Corpus research at Michigan is centred upon that university's English Language Institute. A major contribution of this group has been the construction of MICASE (the Michigan Corpus of Academic Spoken English; Simpson *et al*. 2002), a corpus of nearly 2 million words of transcribed speech occurring in teaching situations at the University of Michigan (seminars, lectures, and so on). The construction of this corpus was initiated by Rita Simpson and John Swales, though many linguists have now worked on the data. This has led to a range of studies on topics including minimal pairs in spoken corpora (Levis and Cortes 2008), verbal stance in spoken discourse (Fortanet 2004) and the use of the corpus to improve the teaching of English for academic purposes (Swales 2002).

The work at Brigham Young (BYU) is particularly noteworthy as it is, in a US context, of long standing. Starting with the work of Randall Jones on German and biblical corpus linguistics (Jones 1997; Jones and Tschirner 2006), BYU has remained a centre of corpus linguistics in the USA. However, activity at BYU has increased in volume and significance in recent years, largely as a result of the pioneering efforts of Mark Davies. Davies has worked to produce, as noted in Chapter 2, a significant fourth-generation corpus search tool at BYU. He has also turned BYU into a major centre for corpus compilation in the USA, building large corpora of contemporary American English, and substantial historical corpora of American English, Spanish and Portuguese.[20] This has not only allowed researchers to explore existing corpora such as the BNC through a new interface; it has also allowed them to undertake diachronic research, using the same interface, on Davies' new corpora (see Davies 2005, 2009a, 2009b, 2010).

4.8 Summary

This chapter has outlined a range of developments in the area of English Corpus Linguistics that have been influential in the development of corpus linguistics in general. While we have focused primarily on a small group of universities where especially pioneering work has taken place, we have also mentioned other institutions and individuals whose contributions have been significant to this field of ECL. In the chapters that follow, areas of study in which corpus linguistics has had a profound impact will be reviewed in depth. In the next two chapters, two topics reviewed briefly here will be returned to and surveyed in much more depth – work on variation in English, including the work of Biber, and the work of the neo-Firthians. Those reviews should be viewed through the lens of this chapter to allow the research presented there to be understood in its broader context. Chapters 7 and 8 move wholly beyond ECL to look in more detail at how corpus linguistics has interacted with, and enhanced, the study of theoretical linguistics and psycholinguistics. Yet even those later chapters should be read with the background of ECL in mind. While the work we discuss there is innovative, many of the basic methods, and some of the ideas, can still trace their roots back to the work on ECL that laid out much of what is today considered the fundamental core of corpus linguistics as a (sub-)discipline.

Further reading

There is little doubt that Aijmer and Altenberg (1991) remains both a very readable collection of papers and a clear overview of the field of ECL as it largely is now, and certainly as it was in the late 1980s and early 1990s. The book is strongly recommended as additional reading to support this chapter. If it lacks anything, it is an account of one crucial area that has come to prominence in ECL since it was published: namely, learner corpora. Readers interested in pursuing this area should consult Granger *et al*. (2002). Another general overview of ECL well worth reading is Meyer (2002), which is a brief and accessible guide to the field. Read in conjunction with Aijmer and Altenberg, it both expands on and refreshes that volume.

Further recommended reading are the books in the Rodopi Language and Computers series. Including, as they do, the outputs of many of the ICAME conferences, these edited collections provide a very useful guide to the development of ECL over time. More recent books published in this series have disseminated outputs from US corpus linguistics conferences as well. So a survey of this book series furnishes the reader with a review not merely of ECL in Europe, but also of the development of ECL in North America.

Further readings for the work of Biber and Sinclair can be found at the end of Chapters 5 and 6, respectively.

Practical activities

(A4-1) Compare and contrast two recent grammars of English which claim to be based on corpus data. Consider the following questions:

- How similar are the descriptions of specific features? (You might pick a verbal feature such as perfect aspect or passive voice, or a feature of the nominal modification system such as adjectives or modifying nouns, or a type of subordinate clause such as complement clause or relative clause.)
- What are the possible sources of difference between the accounts given by the grammars?
- Are there features of one grammar that are absent from the other, e.g. a focus on the grammar of speech? If so, how important, or otherwise, is the distinction introduced?

(A4-2) Search in a spoken corpus for one of the 'filled-pause' vocalisations usually transcribed in English as *um* or *er* (or whatever the equivalent transcription convention is for your corpus). Using the concordance, analyse a random sample of all the sentences in which these typically spoken features occur. Would the grammatical structure of these sentences cause any problems for the type of traditional descriptive grammar that is heavily focused on the written form of the language? If so, to what extent?

(A4-3) Look at a modern language teaching textbook intended for learners of English which claims to be based upon corpus data. What reference does the book make to the corpus data? Is material from the corpus presented directly, or has it been manipulated or organised by the author in some way? To what extent does the use of corpus data allow the book to approach language learning in a way that would not otherwise be possible?

Questions for discussion

(Q4-1) Learner corpora are typically designed to be balanced and representative of a range of L1 backgrounds. However, they may be much less balanced on other criteria of variation – for instance, they are often extremely homogeneous with regard to text type, being composed mostly of essay-style writing. In some cases the corpus builders specify precisely what essay titles must be set when the data is collected! How much of a problem do you think this is for the analysis of learner corpora? What other types of text might you wish to include in a hypothetical 'ideal' learner corpus? Is there a risk that comparability of L1 subcorpora may be lost if the corpus contains different types of text? If so, how might this risk be mitigated or avoided?

(Q4-2) As we have argued in this chapter, English Corpus Linguistics shaped the field of corpus linguistics in general. However, English is hardly typical of all the world's languages! Think of one or more other languages that you

can speak or whose grammar you have studied at any point. How might the development of corpus linguistics have been different, if *that* language had been the hatching-ground of corpus linguistics? Consider what other questions, irrelevant for English, might have been more central coming from the background of this other language; or, alternatively, think about what features of English that attract lots of attention in corpus linguistics are actually odd by cross-linguistic standards.

(Q4-3) When you did activity (A4-2), you should have come up with a list of things that a grammar of the spoken language might need to address. Think now about how you might go about formulating a descriptive grammar to deal with these kinds of things – and, if possible, compare your ideas to the approaches taken in actual speech-oriented grammar research such as Brazil (1995) or McCarthy and Carter (2001).

5 Corpus-based studies of synchronic and diachronic variation

5.1 Introduction

In this chapter, we turn our attention to the issue of linguistic variation, and how corpora have been employed to study differences in the English language across time and across different contexts of language use. We can interpret *variation* in a number of different ways. One is change over time or diachronic variation. In the two sections that follow, we will look at the use of corpora to study language change in pre-contemporary and contemporary English, respectively. Yet while corpus-based analysis of language change is a broad field, the study of synchronic variation is even more extensive. In exploring corpus-based approaches to synchronic variation, we will focus on two rather distinct approaches. One approach, touched on briefly in the previous chapter, is strongly associated with Douglas Biber and colleagues; this is the so-called multi-dimensional (MD) approach. The other is associated with variationist sociolinguistics. Although, as we will see, these approaches have certain commonalities, they are distinct in that the MD approach looks at variation across genre (or register), with the individual *text* as the unit of variation, whereas variationist sociolinguistics looks at variation across class, gender or other social category, with the individual *speaker* as the unit of variation. We will discuss the MD approach, in particular, at some length, because it is methodologically extremely distinct and statistically sophisticated.

5.2 Diachronic change from Old English to Modern English

Looking at language change is an area of linguistics for which corpus data is particularly appropriate. No one now alive speaks Middle English as a native tongue, much less Old English; thus, even if we wish to rely on the judgements of a native speaker, we simply cannot. Instead, for these and other extinct languages there is a fixed 'corpus' of surviving texts which will never grow any further, except in the rare circumstance that hitherto unknown texts are discovered.[1] An electronic corpus composed of all of these surviving texts (or a sampled subset of them) is thus the ideal tool for taking into account as much data

on these historical forms as possible in an analysis of how language has changed. The quantitative analyses enabled by corpus methods are also highly valuable for the study of language change. One quite consistent finding of research in historical linguistics is that one structure very rarely replaces another in a single, sudden change. Rather, new structures arise and are initially used infrequently, and then may later increase in frequency of use, perhaps in competition with some established structure (some examples are discussed in the following section). This kind of quantitative pattern is ideally tracked by a corpus sampling texts across time.

5.2.1 Some notable general historical corpora

In the mid-to-late 1980s, the first major historical corpus of this sort was constructed: the *Diachronic Part of the Helsinki Corpus of English Texts*, more usually known simply as the Helsinki Corpus (Kytö 1996). The Helsinki Corpus' temporal span, reaching as it does from before 850 CE to the eighteenth century, is possibly the widest of any currently available corpus. Furthermore, within the different periods, an attempt was made to cover a variety of types of text, just as would be done in the design of a synchronic corpus.[2] However, since the entire dataset consists of 1.5 million words, the coverage of some periods is inevitably scanty (Kytö and Rissanen 1992: 7–9, 13). Kytö and Rissanen argue, however, that even with this drawback, the corpus allows 'fairly consistent trends of development' to be discerned for many features of English that have changed across this very long period.

Aside from the Helsinki Corpus, one of the most important diachronic corpora to be created in recent years is ARCHER (A Representative Corpus of Historical English Registers; see Biber *et al*. 1993 and Biber *et al*. 1998: 251–3). Like the Helsinki Corpus, ARCHER aims to represent both a spread of time periods and a spread of genres – but the 1.7 million words of ARCHER are focused more narrowly on the last 350 years. Its design has enabled fine-grained analysis of diachronic change from a perspective emphasising genre variation.[3] For example, Hundt (2004a, 2004b; see also Hundt 2007) uses ARCHER, along with other corpora, to look at the relative use of the progressive passive and passival over time. Progressive passive clauses are of the form *X was being done*; the passival is a structure where that same meaning is expressed by a clause which is progressive active in form (*X was doing*): for example *the house is being built* versus *the house is building* (Hundt 2004b: 53). Using ARCHER, Hundt (2004b: 66) shows that 'the increase of normal progressives with inanimate subjects . . . seems to have furthered the near-demise of the passival and the rapid spread of the progressive passive in the second half of the nineteenth century'. ARCHER is ideal for research, such as this, that focuses on the emergence of grammatical structures which especially characterise present-day English. Other research has combined the Helsinki and ARCHER corpora, for example Kytö (1997), who looks at the use of *be* versus *have* as the auxiliary verb in the English perfect aspect

construction. Contemporary English nearly always forms the perfect with *have* (e.g. *he has arrived*); but a perfect with *be* was formerly common with some intransitive verbs (especially verbs of motion; e.g. *he is arrived*), as it still is in other European languages. Kytö uses the combination of ARCHER and Helsinki to conduct a survey of how the frequency of these variant forms of the perfect changes across time and across datasets that is more robust than would be possible using either corpus alone.

5.2.2 Specialised historical corpora

Other historical corpora have been developed to focus on individual text types – for example, the Corpus of Early English Correspondence (Nevalainen and Raumolin-Brunberg 1996), which consists solely of letters, or the even narrower Lancaster Newsbooks Corpus (Hardie and McEnery 2009; Prentice and Hardie 2009), which represents solely the very early news publications of the 1650s. A new trend is a focus on historical *speech* as opposed to writing – a particular concern in the study of language change, as changes often originate in the spoken language and are subsequently transmitted to the more conservative written form. A corpus of actual historical speech would be very limited, given that audio recording does not go back much more than a hundred years (and until relatively recently, no attempt was made to collect samples of spoken language in any systematic way). But we can approach historical speech data by collecting *dialogues*. These are written texts that, on the basis of present-day register variation, are judged likely to approximate speech more closely than other genres of writing (see Culpeper and Kytö 2000). Dialogues, in this sense, include transcripts of court trials and scripts of plays. The Corpus of English Dialogues 1560–1760 assembles over a million words of these types of text (see Kytö and Walker 2006; Culpeper and Kytö 2010).

In summary, then, the range of historical corpora continues to expand to encompass an ever-greater subset of the types of corpora available for present-day English, thus continually increasing the granularity of the historical descriptions and theories that may be derived on the basis of these corpora. However, the analysis of language change over time is not confined to the relatively remote periods of Old, Middle and Early Modern English. The growing field of research discussed in the next section focuses on change across comparatively short periods of time within contemporary or near-contemporary Modern English, particularly in the latter half of the twentieth century.

5.3 Diachronic variation in contemporary Modern English

The work on diachronic corpora outlined above has enhanced our knowledge of change in English through the mediaeval and early modern periods.

Table 5.1 *The Brown Corpus sampling frame*

Text categories		Broad genre	No. texts	% of corpus
A	Press: reportage	Press	44	8.8
B	Press: editorial	Press	27	5.4
C	Press: reviews	Press	17	3.4
D	Religion	General prose	17	3.4
E	Skills, trades and hobbies	General prose	36	7.2
F	Popular lore	General prose	48	9.6
G	Belles lettres, biography, essays	General prose	75	15
H	Miscellaneous (government & other official documents)	General prose	30	6
J	Learned and scientific writings	Learned	80	16
K	General fiction	Fiction	29	5.8
L	Mystery and detective fiction	Fiction	24	4.8
M	Science fiction	Fiction	6	1.2
N	Adventure and western fiction	Fiction	29	5.8
P	Romance and love story	Fiction	29	5.8
R	Humour	Fiction	9	1.8

Note that there are some very slight variations between Brown/Frown and LOB/FLOB in the number of texts per category, as will be apparent if this table is contrasted with Table 1.1.

However, the most recent changes in Modern English have been studied within the framework of corpora originally developed for synchronic analysis using the sampling frame of the Brown Corpus. The Brown Corpus was originally designed as a snapshot corpus covering American English as used in the year 1961 (see section 1.4.3). However, perhaps because of Brown's status as the first widely available electronic corpus, its sampling frame has become a *de facto* standard for the construction of small-scale written corpora – and, of course, as more corpora have been compiled according to this sampling frame, this *de facto* standard has become increasingly entrenched.

5.3.1 The Brown Family of corpora

Compared to later corpora such as the BNC, the Brown Corpus sampling frame is very narrowly specified. A corpus based on this sampling frame must contain five hundred text samples of 2,000 words each (or as close to that length as possible without truncating the final sentence). Furthermore, the texts are distributed in a set ratio across fifteen categories which collectively represent a wide range of genres of published writing. These categories are assigned letters for easy reference. The sampling frame is shown in Table 5.1.

This sampling frame is not without its drawbacks. Some of the text categories and the proportions of the corpus allotted to them may seem somewhat arbitrary. There are also points of difficulty in the description of some of the categories. The term *belles lettres*, for example, is very rarely used as a description of a genre of publication these days, rendering category G somewhat opaque to more recent generations of corpus linguists. Furthermore, the titles of some of the genres can potentially be deceptive. For instance, *Adventure and western fiction* often includes short stories from fantasy fiction magazines and anthologies.[4] To counter some of these problems, it has become usual to group the fifteen lettered categories into the four more easily interpretable 'broad genres' shown in Table 5.1.

The next corpus compiled according to the Brown sampling frame was the Lancaster-Oslo/Bergen Corpus (LOB), collected in the 1970s but sampling British English from the year 1961. Brown and LOB together opened the possibility of analysing synchronic variation between the two most widely used varieties of standard written English. For instance, Hofland and Johansson (1982) undertook an analysis of comparative word frequencies in the two corpora. This line of inquiry was developed further by Leech and Fallon (1992), who analysed and classified the words whose frequency differed significantly between the two corpora, with the aim of developing a picture of the cultural contrast that these differences in lexical frequency imply. They reach the somewhat tentative conclusion that whereas US culture presents itself in 1961 as 'masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise', UK culture appears 'as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life' (Leech and Fallon 1992: 44–5).

However, the potential of these corpora as a resource for diachronic analysis emerged in the 1990s, when a team at the University of Freiburg, led by Christian Mair, compiled a pair of matching corpora for the UK and US English of the 1990s, dubbed Frown and FLOB after their institute of origin and the corpora they replicated. In the meantime, a large number of other corpora had been assembled according to the same or nearly the same sampling frame for other varieties of world English, sampling from the 1970s or 1980s. The result was a set of matching corpora that are known informally as the Brown 'Family' of corpora. As the value of the Family for diachronic work has become clear, work has continued to extend its coverage both forward and backward in time in thirty-year or fifteen-year increments. At time of writing, the core of the Family are Brown, Frown, LOB and FLOB, which allow for simultaneous synchronic and diachronic comparison (see Table 5.2). The more peripheral members of the Family are those corpora which allow only diachronic or only synchronic comparison. However, as the 'gaps' in Table 5.2 are filled in, it is to be anticipated that the core of the Family will extend both before and after its current scope, and to varieties of English other than British and American.

Table 5.2 *The 'Brown Family' of corpora (core members in bold)*

Period covered	American English	British English	Other varieties of English
1900s		Lancaster1901 (in preparation by Leech and Smith)	
1930s	B-Brown (in preparation by Hundt)	Lancaster1931 (aka BLOB, Leech and Smith 2005)	
1960s	**Brown** Corpus (Kučera and Francis 1967)	**LOB** Corpus (Johansson *et al*. 1978)	
1970s			Kolhapur Corpus (Indian English; Shastri *et al*. 1986)
1980s			Australian Corpus of English (Collins and Peters 1988) Wellington Corpus (New Zealand English; Bauer 1993)*
1990s	**Frown** Corpus (Hundt *et al*. 1999)	**FLOB** Corpus (Hundt *et al*. 1998)	
2000s		British English 2006 Corpus (Baker 2009)	

* A corresponding spoken corpus, the Wellington Corpus of Spoken New Zealand English, is also available (Holmes *et al*. 1998).

Matches for Frown/FLOB have also been compiled in other languages (for example, the Lancaster Corpus of Mandarin Chinese: McEnery and Xiao 2004a) as well as some less-close matches such as the *Cronfa Electroneg o Gymraeg* (Welsh; Ellis *et al*. 2001) and the Uppsala Russian Corpus (Lönngren 1993). These additional matches enable cross-linguistic comparison, as noted in section 1.7. The Brown Family is possibly the most comprehensive source of corpus data for studying variation and change in contemporary Modern English. It does have notable disadvantages, however. One is the difficulty experienced at times in applying the sampling frame to a new variety of English or a new language. In the case of both Indian English and Chinese, for example, it proved virtually impossible to find texts to populate some of the categories – few if any stories are written about the American 'wild west' in either Indian English or Chinese. Where such lacunae in the sampling frame were encountered, the sampling frame was adapted to allow such data to be gathered. For example,

McEnery and Xiao (2004a: 1176) describe how they used 'martial arts' fiction texts to populate the N category in Chinese in lieu of 'wild west' adventures, justifying their choice of substitute by demonstrating that it was a similarly popular genre of adventure fiction with a distinct register. However, the sampling frame, with infrequent exceptions such as these, has actually proved to be quite robust for synchronic and diachronic comparison. The second drawback of the sampling frame is, however, much more fundamental. It is partial. There are text types it does not collect – and, equally, it will not allow for distinct new genres, such as blogs, tweets or other genres on the Internet, to be collected. This is excusable as it at least allows comparisons to focus upon a core of genres that are relatively stable over time. The omission of spoken data from the sampling frame is a more fundamental issue, one which undoubtedly limits the value of the Brown Family of corpora as they stand. Some research has worked around this limitation by adding spoken corpora alongside the Brown Family; for instance, Nokkonen (2006) uses the London-Lund Corpus and COLT (Corpus of London Teenage Language) alongside LOB and FLOB in a study of the semantics of semi-modal *need to*. But given that these separate corpora have been shaped by different design goals, this may be a less than optimal way to add spoken data to the analysis.

Another family of matched corpora, the International Corpus of English (ICE; see section 4.2), does include spoken data; it also has the advantage of covering areas such as Hong Kong where English is spoken mainly as a foreign language (what Kachru 1986 dubs the 'expanding circle' of world English), whereas the Brown Family only covers English as a first language (the USA, UK, etc.; Kachru's 'inner circle') or a second language (e.g. India; Kachru's 'outer circle'). However, the ICE datasets lack the diachronic dimension of the Brown Family.

5.3.2 Results from the analysis of the Brown Family

By way of an illustration of the potential of datasets like those within the Brown Family of corpora, in this section we will review briefly the results of exploiting the Brown Family for diachronic analysis. To date, those engaged in the construction and use of these corpora have examined the changes in frequency of a wide selection of grammatical features across different members of the family. These range from the very specific to the highly general. An example of a study exploring a single, detailed feature both synchronically and diachronically is McEnery and Xiao's (2005b) investigation of the usage of the verb *help* in British and American English. By contrast, Mair *et al*. (2002) look at one of the most general of all quantitative measures of grammars, namely the frequency of different parts of speech. The relative frequency of, in particular, nouns and verbs has long been recognised as a key feature differentiating speech and writing, and also different genres of speech and writing (see also, among others, Hudson 1994; Rayson *et al*. 1997, Granger and Rayson 1998, Biber *et al*. 1999: 65, 235, *passim*;

and Rayson *et al*. 2002). Mair *et al*. find some evidence that these frequencies also vary diachronically between 1961 and 1991 – so, for example, in each of the four broad genres of FLOB, there are more words tagged as nouns than in the equivalent genre of LOB. This may be taken as indicating a shift over time to a more 'nominalised' style. The frequency of verbs, however – which are typically more common in speech – is more stable over time.

A much wider set of grammatical features, and their distribution across all four core members of the Brown Family, are surveyed by Leech and Smith (2006), who furthermore suggest a set of reasons for the changes observed. For many of the grammatical features they investigate, they argue that the change is in the direction of writing becoming more like speech. That is, features that are synchronically more common in speech than writing are, over time, coming to occur more often in writing. Examples of this include the auxiliary and negative contractions (*I'm*, *she'll*, *isn't*, *won't*). Contrariwise, features that are synchronically more common in writing than speech are coming to occur *less* often in writing (for example, the frequency of the passive construction is declining over time in written English). These changes are summarised as forms of *colloquialisation*, 'a tendency for features of the conversational spoken language to infiltrate and spread in the written language' (Leech 2004: 75). But there is also a trend of *Americanisation* in British English – that is, a tendency for British English to follow changes in American English. For example, a type of subjunctive construction called the mandative subjunctive,[5] particularly characteristic of American English, appears to have become more common in British English between 1961 and 1991, though it remained in 1991 less common than it is in American English. This is despite the fact that the mandative subjunctive is decidedly non-colloquial. Colloquialisation and Americanisation, then, can motivate grammatical change but do not necessarily push it in the same direction.

Some of the grammatical features included within Leech and Smith's discussion have been examined, individually, in greater detail. For instance, Smith and Rayson (2007) focus in depth on the progressive aspect, and in particular the progressive passive. As we noted above, Hundt (2004a) tracked the rise of this construction in the Modern English period; Smith and Rayson investigate its fortunes between 1961 and 1991. They conclude that it is likely that the frequency of use of the progressive passive continued to increase in British English over this period – and, furthermore, that the passival has all but vanished. However, of the many areas of recent grammatical change that have been investigated, the one that has attracted the most scrutiny is perhaps modal verbs.

The frequencies of the central modals and semi-modals are among the features examined across the core Brown Family by Leech and Smith (2006); Leech (2004) extends this analysis to spoken data from the SEU and ICE. Let us first outline the general picture which emerges from a study of modal verbs across the Brown Family. The broadest finding is that the use of modal verbs declined significantly between 1961 and 1991, in both British and American English. Furthermore, the observed decline is most severe for those modal verbs, such as

shall, which were relatively rare to begin with. However, the declining modals are not being directly replaced by equivalent semi-modal constructions; some semi-modals have become more common (e.g. *want to*) whereas others have become rarer (e.g. *be to*) (see Leech *et al*. 2009: 97). Moreover, notably, while the decline in use of modals is most advanced in American English, it is in *British* English that semi-modals are most common (Leech 2004: 65–70). So it seems that Americanisation may be a factor in explaining what is happening to the modals – but not the semi-modals. The other factor previously mentioned, colloquialisation, may be important for the semi-modals, but Leech also suggests two other relevant trends. The first is *grammaticalisation* (Hopper and Traugott 1993), the process whereby, over a period of centuries, constructions based on lexical words become bleached of their semantic content and begin to behave like grammatical constructions. This process is, for instance, responsible for the derivation of the semi-modal *be going to* marking future time from the lexical verb *go* meaning 'move'. Whether changes in frequency over much shorter periods of time can be ascribed to grammaticalisation is less clear. The final trend which Leech cites is *democratisation*, the social tendency for people to avoid 'unequal and face-threatening' ways of expressing a particular meaning; that is, a move away from explicit linguistic marking of social power relations. This may explain changes such as the decrease in the use of (deontic) *must*, which implies the exertion of control by one individual over another, and the rise in the use of the near-equivalent semi-modals *have to* and *need to* (Leech 2004b: 75–6).

So the forces for diachronic change – Americanisation, grammaticalisation, colloquialisation and democratisation – may push in contrary directions, operate on different timescales and have different motivations (grammaticalisation being a linguistic trend, for instance, whereas Americanisation and democratisation are purely sociocultural in nature). As Leech says, we find these trends expressed *patchily* in the overall suite of grammatical changes observable in recent English; different changes – not only in modal verb usage but in all areas of the grammar – are attributable to different (combinations of) trends.

Leech notes as a reservation (2004: 70–1) that these observations and speculations rest on the assumption that the Brown Family is sufficiently representative of British and American Standard English in 1961 and 1991 (in the technical sense of *representative*: see sections 1.4.3 and 1.4.4). It is, of course, in principle not possible to determine the representativeness of the Family empirically – since we can never compare the Family to the full population of texts that it samples to determine how accurately it reflects their nature. However, Leech argues that given the high statistical significance of many of the observed frequency changes, and the replicability of the diachronic patterns across different grammatical features and across the four broad genres within the Family, the results may be deemed plausible, although the representativeness issue prevents us from considering them definitive. This argument is mostly persuasive. However, other data presented by Millar (2009) which also bears on recent change in the use of modal verbs may suggest the picture provided by the Brown Family is not

wholly unproblematic. Millar examines modal verb frequency in each year of the *TIME* magazine corpus.[6] The time-lapses in this analysis are thus much shorter than the thirty-year gap between 1961 and 1991. Furthermore, although the corpus is clearly much less representative of written English as a whole than the Brown Family, it is much *more* representative of the sub-population of texts that it samples – because it is in effect a 100 per cent sample of a narrowly defined text type, i.e. *TIME* magazine. One of the most surprising results that emerges from this analysis is that, although there *are* trends over time in *TIME*, there is *dramatic* variation in the frequency of particular modals from year to year within the corpus. For some modal verbs, the year-to-year fluctuations are actually as big as the thirty-year changes. This problematises the practice in diachronic studies of using the Family to, in effect, plot a point on a graph for 1961, plot another for 1991 and then draw a straight line between them. If the variation in texts drawn from a single publication may be so drastic from year to year, can we maintain confidence in the statistical significance ascribed to the thirty-year differences in the Family on the basis of their very large relative size? There is, in addition, some evidence that the frequency of even such core grammatical features as modal verbs may be responsive to relatively short-term sociohistorical pressures. Millar's data reveals, for instance, that the patterns of modal usage are noticeably different in issues of *TIME* magazine published during the Second World War!

On the one hand, the methodological issues raised by Millar's data are substantial and must, at some point, be addressed directly. On the other hand, the *replication* of results in diachronic analyses of modal verbs is very impressive – not only for different genres within the core Brown Family, but for comparable analyses undertaken on peripheral members of the Family or on other corpora altogether – and not easily explicable on the assumption that the measured differences are in effect random artefacts.

Let us return to the ARCHER corpus to illustrate this. Biber (2004; see also Biber *et al*. 1998: 205–10) used ARCHER's ten register categories to examine the frequency of modal verbs across time *and* across registers. His main conclusions are that, from 1650 to the present, 'modals have been decreasing in frequency in all registers, but semi-modals have shown noteworthy increases only in drama and personal letters', but that 'the patterns of change for individual verbs often differ across registers' (Biber 2004: 210). These findings are decidedly compatible with Leech's observations: modals are in decline but *not* because of a mirror-image rise in the use of semi-modals. Biber extends this analysis to grammatical markers of stance apart from the modals and semi-modals, such as stance adverbs and stance complement clauses. He concludes that many devices for the expression of stance have increased in frequency over time, and that speakers of English have been in general more willing to express stance explicitly in more recent historical periods. Again, this sociohistorical analysis is not incongruent with Leech's thoughts on colloquialisation and democratisation. If we embraced the most radical interpretation of Millar's year-to-year data – that measurements made on snapshot corpora separated by periods of several decades are fundamentally

spurious – we would be hard-pressed to explain why similar pictures of diachronic change emerge from different corpora, since on that assumption we would expect the results from different corpora to differ randomly. But the pictures emerging from studies of the Family and studies of ARCHER clearly *are* similar. Thus, a more cautious interpretation of Millar's data is called for. We might posit, perhaps, that the use of a sampling frame such as that of the Brown Corpus somehow evens out the large year-to-year differences observable in a single-source corpus such as *TIME*. But it remains to be worked out in detail exactly how this would come about. In the meantime, we would argue that in extending the Brown Family to the twenty-first century, snapshot periods shorter than three decades may be prudent, where practical. In fact, exactly this approach has been taken in the construction of British English 2006 (Baker 2009), a match for LOB that samples 2005–2007 – a lapse of *fifteen* years from FLOB, not thirty.

As we noted, Biber's (2004) study of modality and other markers of stance over time was also a study of registers. *Registers* are groupings of texts defined by external factors – that is, social or situational features of the medium they use, the context in which they were produced, or the purpose of their creation. It is in fact his approach to variation across registers which forms the core of Biber's research methods, and it is to this approach to the study of language that we now turn.

5.4 The multi-dimensional approach to variation

5.4.1 An overview of the MD method

The multi-dimensional (MD)[7] approach to studying textual variation is associated primarily with Douglas Biber. It was first introduced in a study (Biber 1986) which aimed to explain certain puzzling findings in earlier work on variation between speech and writing, and between different registers of each. Biber (1986: 386) argues that work in this field had produced contradictory results because of differences in the *linguistic features* taken into account when contrasting two or more registers. Biber suggested an innovative, much more comprehensive approach, which looks at the use of a very large range of features of language in different registers and uses statistical techniques to weave them together into a more complicated and subtle picture of how registers differ from one another.

This more sophisticated approach looks at a *list* of sixty-seven linguistic features, in contrast to earlier studies which typically focused on one feature or a much smaller group of features (hence it can be labelled a 'multi-feature' approach). Biber developed this list of features on the basis of a survey of the previous literature on distinguishing spoken and written discourse. The next step in the MD method is to measure the frequency of each of the features within

a corpus sampled from a heterogeneous set of registers. Biber's (1988) study of Standard English used samples drawn largely from LOB, covering all fifteen sections of the Brown sampling frame, but also adding various spoken texts from the London-Lund Corpus (see section 4.2) as well as samples of personal and professional letter-writing. Subsequent MD studies (discussed below) have employed corpora of comparable size and heterogeneity.

A statistical analysis is then applied to (normalised) frequencies of the many linguistic features. The purpose of this statistical procedure – called a *factor analysis* – is to cluster together linguistic features which tend to vary with one another.[8] This is summarised by Biber *et al*. (1998: 278) as follows:

> In a factor analysis, the correlations between a large number of variables (i.e. the linguistic features) are identified, and the variables that are distributed in similar ways are grouped together. Each group of variables is a factor – which is then interpreted functionally as a dimension of variation . . .

So, for instance, when Biber (1986, 1988) applied this analysis, it emerged that texts which contain (relatively) many past-tense verbs *also* tend to contain (relatively) many third person pronouns. So these two features are grouped together. The factor analysis continues in this way until it has reduced the very large list of linguistic features to a much smaller number of *factors* that describe the variation among the texts in the dataset. The key to the MD approach, and the reason why it is 'multi-dimensional', is that these factors are interpreted as *dimensions* – that is, independent scales along which a text can vary (and as there are many factors, the approach is multi-dimensional). For instance, we know that a text might be formal or informal. We also know that a text might concern concrete subject matter or abstract subject matter. However, there is no necessary relationship between these two parameters of the register. A text on abstract matters could equally well be formal or informal, as could a text on concrete matters. More subtly, there is also the possibility of *gradience* – rather than a binary opposition – between concrete and abstract, formal and informal. So formality and concreteness of subject matter are independent, and we can imagine them as the x and y axes on a two-dimensional graph. Variation in abstractness might affect the 'horizontal' position of a text on the graph, variation in formality the 'vertical' position – that is, two independent sliding scales.

The visual metaphor of two dimensions for two independent forms of register variation is an appealing and powerful one. However, it is problematic because choosing abstractness and formality as our two dimensions is essentially arbitrary. Why only two types of variation? Why not three or four? Where do you stop adding dimensions? Which dimensions of variation in situation, context and purpose are actually significant for studying variation in *language*? Biber's MD method offers a way around these questions. The factors which emerge from the statistical analysis discussed above must, by virtue of the way they have been calculated from linguistic feature frequency statistics, necessarily represent aspects of variation that *are* significant for studying language. So Biber argues

that the factors can be considered as *dimensions* of register variation. On the basis of the linguistic features grouped together on each dimension, Biber proposes a functional interpretation for that dimension. The dimensions thus link together the functional requirements of a particular register with particular linguistic features that are favoured by those functional requirements.

This is quite hard to understand in the abstract, so let us take a particular example. Factor 2 in Biber's statistical results puts together the following linguistic features (Biber 1988: 102, 108–9):

- *high frequency of*: past-tense verbs, third person pronouns, perfect aspect verbs, public verbs, synthetic negation, present participial clauses;
- *low frequency of*: present-tense verbs, attributive adjectives.

Biber argues that features such as the past tense (and, thus, relative lack of present tense) and third person pronouns relate to the function of narrative discourse: namely, the relation of *past events* with *specific participants*. Meanwhile, a high frequency of attributive adjectives is associated with elaborate noun phrases, a feature of non-narrative discourse, and so the opposite feature – *low* frequency of attributive adjectives – becomes associated with narrative. Biber thus allots to this dimension the functional label 'Narrative versus Non-Narrative Concerns'.

Biber (1988) proposes six such dimensions from seven statistical factors, one of the factors being statistically too weak to be safely interpreted as a dimension. An earlier version of his analysis of Standard English (Biber 1986) found three dimensions; in later work (e.g. Biber 1989), he tends to focus on only the five strongest factors, to which he assigns the following functional labels:

- Dimension 1: 'Involved versus Informational Production'
- Dimension 2: 'Narrative versus Non-Narrative Concerns'
- Dimension 3: 'Explicit versus Situation-Dependent Reference'
- Dimension 4: 'Overt Expression of Persuasion'
- Dimension 5: 'Abstract versus Non-Abstract Information'

Biber does not stop at identifying these dimensions but uses the raw statistics for the linguistic features to produce an overall 'score' for each register on each dimension. As noted above, these vary incrementally, so on each dimension there are registers with high scores, registers with low scores and registers with middling scores. Furthermore, because the dimensions are substantially independent, the relative ordering of the registers may be completely different from dimension to dimension. The registers that score low on Dimension 2, for instance, include telephone conversations, professional letters, academic prose and official documents. The registers that score high include all kinds of fiction, biography and spontaneous speeches (Biber 1988: 136). This precise spread is not observed on any other dimension. An illustration of the spread of registers on Dimension 2 is given as Figure 5.1.

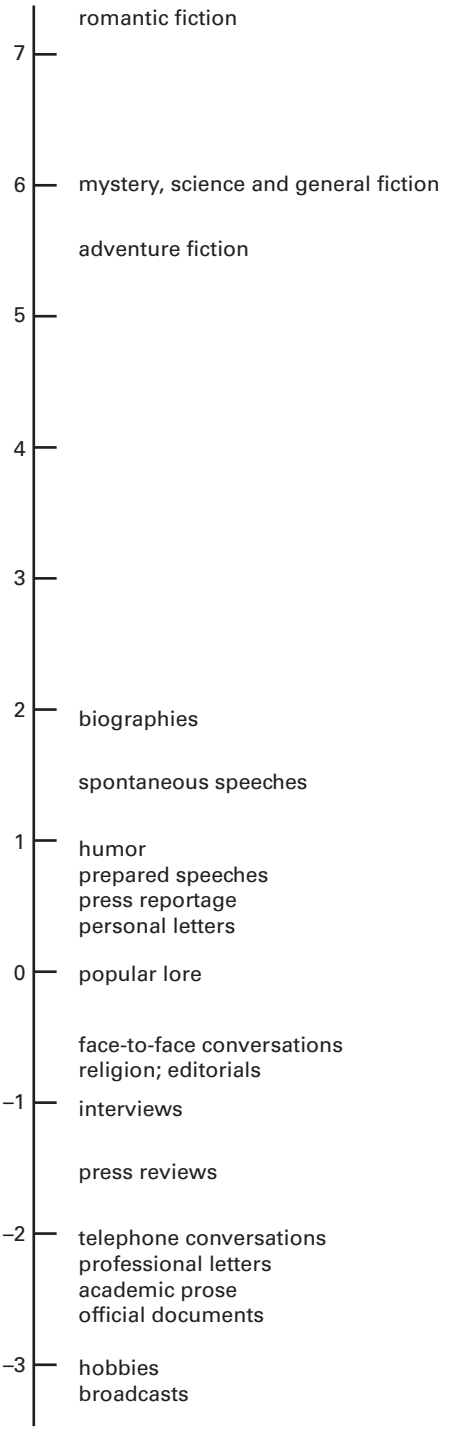

Figure 5.1 *Biber's Dimension 2,* Narrative versus Non-Narrative Concerns*; narrative texts have high scores, non-narrative texts have low scores. Reproduced from Biber (1988).*

Biber (1986) argues, critically, that no one dimension equates to a straightforward distinction between speech and writing, and that all the dimensions are needed to characterise the differences between the registers in his dataset. We can see this from the functional labels on the dimensions. Prototypical conversation is involved rather than informational – but this is not true of *all* spoken registers: consider broadcast speech such as, say, a television news report. Similarly, prototypical conversation tends to make use of situation-dependent reference rather than explicitly spelling out what is being discussed – but some spoken registers do *not* do this, such as prepared and spontaneous speeches. In fact, the oral–literate distinction is represented to some degree by three different dimensions – 1, 3 and 5. Equally, on Dimension 2 (see Figure 5.1), there are both spoken and written registers that score very highly, as well as both spoken and written registers that score very low.

It is in this way that Biber uses the results of his MD approach to English to explain the previous contradictory findings of research in the 1960s, 1970s and early 1980s. Because only one or a few linguistic features were being investigated to map out the difference between speech and writing, earlier researchers could not identify multiple dimensions of variation. Thus, Biber's initial work on the MD model resulted in the most detailed model of register variation yet proposed, providing a notable demonstration of the power of radically corpus-based, computational and statistical approaches to problems in text and language analysis.

5.4.2 Applications of the MD approach

Since developing the MD approach, a clear strand of Biber's research has focused on extending the method or applying it to new areas of investigation. The pattern of register variation developed in Biber (1988) has proven a useful starting point for a number of diverse analyses. For example, Biber and Finegan (1989) undertake a diachronic analysis within the MD framework of Biber's (1988) dimensions. Using a dataset of fiction, essays and personal letters from the past four hundred years, Biber and Finegan examine the evolution of style within these three registers, using the three dimensions linked to the oral–literate distinction as a basis for comparison. A key methodological point is that Biber and Finegan do not repeat the process that produced the dimensions on this new data – as that would not allow comparison between datasets. Rather, the dimensions established on the basis of Biber's (1988) corpus are treated as given, and the historical genres are positioned on those dimensions. Biber and Finegan find major changes over time relative to each of the three dimensions; furthermore, the patterns for each register on each of the three dimensions are generally alike: '17th- and 18th-century texts tend to be moderately or extremely literate, with a transition towards more oral styles in the 19th century and the development of a distinctly oral characterisation in the modern period . . . across the four centuries

all genres have tended towards more involved, more situated, and less abstract styles' (Biber and Finegan 1989: 507). They convincingly connect this finding to the social history of literacy and style across this period. In contrast to Biber's (2004) work on modals over time, and indeed to most diachronic corpus analysis, Biber and Finegan's emphasis is very clearly on the development of the registers, rather than of any particular linguistic feature or features.

In another direction, Biber (1989) applies the MD approach to the study of text types. Text types are distinct from registers in a crucial respect: while a register is a group of texts defined on the basis of language-external features (i.e. context, medium, purpose), a text type is a group of texts defined on the basis of linguistic similarity, with no necessary implication that they are from similar contexts of use. Biber again uses the dimensions previously established, this time as a means of measuring the distance of each text in his data from each other. On the basis of these distances – relying on all five dimensions, and thus on a very large proportion of his underlying list of linguistic features – he uses a *clustering* procedure (rather than a factor analysis on this occasion) to group texts into eight distinct text types.[9] These text clusters form groups in some ways similar to the registers. However, Biber argues (1989: 39) that this approach identifies finer distinctions among text types than were previously considered to exist; for instance, Biber identifies two different text types with characteristics of narrative, whereas earlier studies had often described 'narrative' as a unitary text type. Of course, given that narrativity is a *dimension* in Biber's model, it is not surprising that a high position on that dimension should characterise more than one text type.

Perhaps the most surprising application of the MD approach has been as a tool for cross-linguistic analysis. Biber (1995a) presents the results of a lengthy programme of collaborative research applying MD methods to corpora of Somali, Korean and Nukulaelae Tuvaluan – three languages genetically and geographically quite separate from one another and from English. In view of this, it is interesting that, when the dimensions that emerge for each language are compared, the overall picture is one of *similarity* (Biber 1995a: 278). All three of these languages, like English, have multiple dimensions that relate to the oral–literate distinction; all lack a single, clear speech-versus-writing distinction. Furthermore, 'each language has dimensions that mark personal stance [. . . and . . .] a dimension marking narrative versus non-narrative discourse' (Biber 1995a: 237). These findings are potentially highly significant, for functional and typological linguistics as well as text linguistics and corpus linguistics, in that they may point the way towards universals of the *functions to which language is put* – which, in turn, affect linguistic choices in ways captured by the different dimensions.[10] More recently, the MD method has been applied to Spanish (Biber *et al*. 2006), with the same kinds of pattern emerging.

Biber's most recent work within the MD framework has focused on language as it is used in the context of universities. The ultimate goal of this undertaking is to assist, especially, non-native speakers in English-speaking higher educational

settings (Biber 2006: 2). Using a tailor-made corpus of 2.7 million words, and a list of 129 linguistic features, Biber undertakes a full MD analysis of the different university registers (textbooks, academic speech in different contexts and so on). In this case, then, the prior results of Biber (1988) did not form part of the analysis. Biber (2006: 181–2) argues that this procedure is appropriate when approaching a new discourse domain, where that domain contains many registers. Another way to think about this might be that the registers represented within the university domain only overlap marginally with the domain of general English examined by Biber (1988) (for instance, the register of academic writing is present in both). It cannot therefore be assumed that the same dimensions will differentiate registers in the new domain. Some functional factors that were relevant to general English may not apply; likewise, university-specific functional factors may emerge. In fact, three out of the four dimensions that emerge for the university registers have parallels to a dimension identified for general English (Biber 2006: 211).

The MD approach is not the only methodology applied in Biber's study of university language. Vocabulary usage, variation of particular grammatical features and the expression of stance by a variety of linguistic devices all form a part of his analysis. In every case, the variation of these features *across separate registers* in the university domain is the key focus of Biber's analysis. Another major element of the analysis is *lexical bundles* – that is, highly frequent multi-word sequences such as *in the light of*.[11] The notion of lexical bundles was first used in Biber *et al*.'s (1999; Chapter 13) corpus-based reference grammar of English (see also Biber and Conrad 1999).

Methodologically and technically, 'lexical bundles' are simply n-grams – recurring sequences of *n* words. However, the term has come to be associated with Biber and colleagues' particular approach to the interpretation of n-grams, rooted in their descriptive goal of capturing the overall characteristics of registers (individually or in contrast to one another). This approach, most notably, concentrates on the 'structural and functional' interpretation of the lexical bundles (Biber 2006: 172). As with the dimensions in the MD method, the aim is to explain *functionally* the frequencies of particular bundles across registers. Biber also notes that '[t]he patterns of use for lexical bundles are strikingly different from those found for traditional lexico-grammatical features'. For example, consider the lexical bundles identified by Biber *et al*. (2004: 384, 389–90) as having the function in classroom teaching of expressing *personal epistemic stance*: these include *I don't know if*, *I don't know how*, *you know what I*, *I thought it was* and others. Although they have various grammatical structures, these all work to express uncertainty in some way – and share the fixity of form that results in their identification as lexical bundles.

In sum, then, Biber's applications of his MD model have contributed significantly to many different subdisciplines of linguistics. However, the approach has not been more widely adopted. In a generally positive review of Biber (1995a), Kilgarriff (1995: 613) notes that:

> [t]he MD research program has been proceeding for over a decade, but as yet the methodology has only been used by Biber and a small group of collaborators. This could be because other readers have not been impressed by the work, or it could just be that the methodology is technically difficult and time-consuming to implement. I suspect the latter. Each stage of the methodology involves a new set of obstacles and skills, and an MD analysis is not lightly undertaken.

Regrettably, this is still true, more or less, at the time of writing – sixteen years on from Kilgarriff's review and twenty-five years on from the initial appearance of Biber's (1986) paper in the journal *Language*. Other methodologies developed by Biber and his colleagues have been picked up and exploited to full advantage by other researchers. For example, Culpeper and Kytö (2002) investigate the use of lexical bundles in dialogue text in Early Modern English. Lexical bundles, however, are computationally and statistically much more straightforward than the factor analysis that underlies an MD analysis. It is almost certainly this complexity that has inhibited the widespread uptake of what appears to be a useful technique. It has been argued by Tribble (1999) and by Xiao and McEnery (2005) that the methodological complexity of the MD analysis can be reduced by using a keywords analysis to achieve much the same effect; likewise Crossley and Louwerse (2007) show that an MD analysis is possible using only bigram frequencies as the input features. However, that aside, the lack of uptake of MD methods by a wider community supports, we would argue, a point we made in section 2.5.4: it is imperative that corpus tools accessible to the non-technical user should continue to develop and to incorporate more and more specialised and complex procedures, up to and beyond the level of complexity of an MD analysis. Unless this occurs, and procedures like MD analysis become as easy for the user to run as a straightforward concordance, without technical training linguists will never be able to access and benefit from the full gamut of corpus methodologies.

5.4.3 Criticisms of Biber's MD methodology

While we have discussed briefly a practical impediment to the uptake of the MD approach, Biber's approach to language variation has also been subject to more fundamental criticism. Not all such criticisms have been entirely well founded, however. For example, Watson (1994) attempts to apply Biber's MD model to the analysis of one postmodern author's novels, a purpose quite different from the model's original use, and on the basis of problems encountered during this effort argues for a set of deficiencies in the MD approach per se. The publication of this paper sparked a debate in the pages of the journal *Text* (Biber 1995b; Watson 1995) in which Biber refutes Watson's criticisms comprehensively. This debate is in itself of limited significance, but it serves to underline the critical importance of careful methodological awareness in the application of corpus linguistic techniques.

Three criticisms of the MD approach, however, deserve some further consideration. All are broadly methodological. The first – and least serious – is the nature of the dataset on which Biber's study of spoken and written English was based (and which thus was also the foundation of several of Biber's later investigations). It is rather small and consists of short samples of longer texts (like LOB, from which it was largely drawn). In fact, we might say that Biber's MD studies, especially those of languages other than English, epitomise the small, carefully designed sample corpus approach pioneered by UCL and Lancaster (see section 1.4.3), and thus the drawbacks of that kind of dataset necessarily apply in full to the MD framework. However, in response to criticisms along these lines, Biber has done important empirical work assessing the impact of corpus representativeness (Biber 1990, 1993). For example, Biber (1990) demonstrates that re-running his MD analyses on small subsets of his corpus produces very similar results to the original analysis. He argues that this supports the contention that the corpus is sufficiently representative.

A more important criticism concerns the replicability of Biber's results. While much of Biber's own further research supports his original results for English, Doyle (2005: 4) points out that others have had difficulty replicating Biber's findings independently, due in part to the unavailability of relevant software and datasets. Furthermore, there are reports in the literature of data suggesting that some of Biber's dimensions may not be statistically replicable in other general English datasets (Lee 2001). If we were to take a devil's advocate approach to Biber's (2006) MD analysis of university language, which found similar but not identical dimensions to Biber (1988), we might argue that this actually demonstrates that the dimensions that emerge are relative to the spectrum of particular texts under analysis at any given time. Hence, no strong claim can be made for the validity of dimensions beyond the corpus they are calculated for. This may be a legitimate argument as far as the *precise identity* of dimensions is concerned, but the full body of Biber's (cross-domain and cross-linguistic) MD research provides powerful support for consistency in the *general outline* of observable dimensions. Replicability remains, however, something of a concern for the MD framework.

Finally, some criticism has been directed at Biber's approach to the choice of linguistic features to use in order to generate the corpus statistics that drive the factor analysis calculation. Altenberg (1989: 171–3) points out that the choice of features is actually one of the keys to the results of a factor analysis, and that some of Biber's (1988) features are questionable on one ground or another. For example, the choice of features is limited to what can be searched for in a part-of-speech-tagged, but not parsed, corpus. Developing this criticism, Ball (1994) discusses in detail some of the problems that can arise in searching corpora for grammatical patterns, and argues that studies like Biber's, based on multiple features whose identification by a corpus search is not unproblematic, are premature. A different kind of problem emerges with regard to some other features. Altenberg (1989: 172) points out that preposition phrases (which constitute just one of Biber's sixty-seven features) are functionally heterogeneous, as they can either postmodify a

head noun or mark an argument or adjunct of a verb. Because they are counted as a single feature, however, the MD method cannot possibly distinguish between these two functions of preposition phrases. But there is no particular reason to assume a priori that the two functions have the same pattern of variation across registers or across texts. For example, if some register has a close-to-average frequency of preposition phrases, we cannot know whether this is because both sorts of preposition phrases occur with middling frequency, or because one sort is very frequent and the other sort is very rare – since both these cases would lead to a middling frequency on average.

This point also emerges within Biber's own multilingual work. One of the preliminaries to the MD analyses in Biber (1995a) is an analysis of Somali versus English registers at the level of frequency of some set of *individual* linguistic features. One of the major findings of this exercise is that establishing clear correspondences between individual grammatical features in different languages is extremely difficult (Biber 1995a: 73–82). For example, English and Somali both have relative clauses – but while they are structurally similar, they are vastly more frequent in Somali, suggesting that 'they are serving very different functions in the two languages' (Biber 1995a: 75). A similar argument applies to English and Somali's marking of oblique verbal arguments such as indirect objects – namely, prepositions in English and preverbal case particles in Somali. Biber observes, just as Altenberg does, that English preposition phrases *also* have the function of nominal modification; but this is not the situation in Somali, where preverbal case particles function solely as part of the verbal modification system. Biber notes, in fact, that English noun-modifying preposition phrases are more similar, functionally, to Somali relative clauses (which also modify nouns). Furthermore, if the different English and Somali features are aligned into two groups, those that modify verbs and those that modify nouns, regardless of the structural details, then the cross-register frequency variations in nominal modifiers and verbal modifiers are quite similar across English and Somali (Biber 1995a: 79–80). Biber argues that because individual linguistic structures may not be well defined functionally, a cross-linguistic comparison of *dimensions* (subsequent to an MD analysis) is better than comparing small sets of particular features when investigating register variation in multiple languages, because the dimensions are 'by definition the functionally important linguistic parameters of variation represented structurally in each language' (Biber 1995a: 82–4). Yet if individual features, such as English preposition phrases, are acknowledged to have multiple functions – corresponding to two or more features in Somali that may quite possibly be linked to different dimensions in a factor analysis – we may be justified in withholding full confidence in the dimensions derived from patterns of usage of those individual features.

Criticisms such as these are bound to arise for two reasons; firstly, as Altenberg (1989) notes, '[t]he classification of linguistic forms into functionally clearcut categories is notoriously difficult'; but also because of Biber's procedure for choosing linguistics features, which is fundamentally ad hoc. For English it is based on literature survey, and in his analyses of other languages it arises from

- 'Meaning'
  - Actions (verbs)
    - Auxiliary forms
      - Aspect
      - Voice
        - Passives with by-phrase
        - Passives without by-phrase
      - Modality
    - Verb forms
      - Past
      - Present
    - Verb modification
  - Entities (nouns)
    - Noun forms
      - Proper nouns
      - Pronouns
      - Derived nouns (nominalisations)
      - Plurals
    - Noun modification
      - Adjectives
      - Articles
      - ...

Figure 5.2 *A fragment of a feature tree for English.*

'existing grammatical descriptions, exploratory analyses of texts from different registers, and analogy to functionally relevant features in English' (Biber 1995a: 94). But there is no mechanism in this approach that can prevent potentially relevant features from being missed. As Biber's own discussion of Somali case particles makes clear, it is important to take into account *all* of the features making up a more general functional category (such as noun modification or verb modification), as well as considering individual features at a low level of generality. So we might argue that the MD methodology could be more solidly founded if based on a selection of features which is both *principled* and *exhaustive* – a standard which Biber's feature-lists approach but do not reach. How might such a fully motivated list of lexicogrammatical features be derived? At this point we move into the realm of hypothesis. However, one possible approach is to consider the functions of a language as a feature tree. This could start at the very high level of nominal components versus verbal components (since the noun–verb distinction is one of the most universal features of language structure), and then diversify from there, with attention to contrasting linguistic options and category alternatives at each branch in the tree. Figure 5.2 illustrates part of what such a feature tree, rooted at the most absolutely general linguistic function (expression of 'meaning' in the very broadest sense), and containing several of the features Biber in fact includes in his MD model for English, might look like.

It would be conceptually straightforward, if practically extremely challenging, to develop an ordered tree that could encompass all of the features used by Biber, that would be consistent between languages at the upper levels, and would make it obvious where features have been overlooked. Such a tree would probably end

up including most of the grammar of the language in outline form – but, surely, a comprehensive feature set should be the ideal starting point for the MD method.

Such speculation aside, there are, in summary, questions to be raised about the systematic choice of linguistic features for an MD analysis, and in addition there is reason to suspect that the choice of linguistic features may affect the outcome of the factor analysis. Again, however, the counterargument is that Biber's cross-domain and cross-linguistic MD analyses, which *all* exploit different feature sets to his (1988) study, produce sets of dimensions that are compatible with, though not identical to, those observed in general English. So if the feature set's makeup does have an effect on the dimensions that emerge, it cannot be significant enough to alter completely the outcome of the MD analysis.

5.4.4 The MD approach: a summary

There is little doubt that the MD approach has proven influential. But how is its impact upon the study of language to date best summarised? On the whole it has shifted debate rather than reframed it. On the one hand, the MD approach has led linguists away from easy dichotomies of spoken versus written language which do not capture the subtleties of the variation of language in use. On the other hand, Biber's various MD analyses have typically produced, as one of their primary results, an oral-versus-literate dimension or dimensions (Biber 2006: 186; 1995a). This is clearly very similar to the speech–writing continuum often assumed in non-MD contrastive analyses. So while it expands greatly the amount of information and the degree of detail that can be employed in the comparison of registers, the MD methodology should perhaps be regarded as an evolutionary rather than a revolutionary advance.

5.5 Corpora and variationist sociolinguistics

The alert reader may have noticed that the type of 'variation' in language we have been discussing so far is variation across or within registers – that is, where the unit of variation is the individual text. Of course, there are other approaches to variation in linguistics, for example pragmatic variation (Barron and Schneider 2009) and variations in discourse (Jucker *et al*. 1999). One of the most prominent is what is sometimes called *variationist sociolinguistics*, that is, sociolinguistic research in the tradition of Labov (1969, 1972), exemplified by the work of contemporary scholars such as Trudgill, J. Milroy, L. Milroy, Cheshire and Kerswill. This tradition in sociolinguistics is, like Biber and others working on the comparison of registers, interested in synchronic variation; but it is also concerned with diachronic change, as are Leech and colleagues and the other scholars reviewed in the first part of this chapter.

So far, corpus linguistic methodologies have not played an extensive role within variationist sociolinguistics. There are multiple reasons for this. There is, to begin with, the practical issue of phonetic transcription. While sociolinguists are often interested in variation of pronunciation, the phonetic transcriptions necessary to assess this are usually absent from large spoken corpora. Aside from this, the main point of disconnect between sociolinguistics and corpus linguistics is that variationist sociolinguists are typically interested in variation at the level of the individual speaker, rather than at the level of the text. So while 'external' variables relating to the situation in which a discourse was produced are not without interest in sociolinguistics, as in corpus linguistics, sociolinguists also study external variables of speaker identity such as gender, ethnicity and class (see Besnier 1998: 127).

So, for example, in Biber's MD approach, a 2,000-word sample of a spoken conversation might be used as one of the texts exemplifying the register of conversation within a broader sample of registers. However, there would not be any methodological differentiation between the speakers in this sample. The method operates at the level of the text, not the level of the speaker, and the aim is to characterise the language of the register, not the language of a particular group of people. For the most part, this could hardly be otherwise. In the study of written language in particular, corpora have represented for the most part published texts – that is, texts which have gone through an editorial process and thus can no longer be unambiguously identified as the production of a single speaker. Spoken corpora with relatively rich speaker metadata – for example, the demographically sampled part of the spoken BNC (see Crowdy 1995) – take a step towards the sociolinguistic treatment of the speaker as the unit of variation, and studies such as Schmid (2003) and Rayson *et al*. (1997) have studied gender variation in this data. But even in the spoken BNC, a perhaps more typical analytic approach is to bundle large numbers of speakers or situations of speech production together. For example, the Rayson *et al*. (2002) study groups together *all* the spoken-demographic data, in contrast to *all* the spoken-context-governed data, without any differentiation as to speaker. Sampson's (2002) study, which looks at the use of the perfect aspect in different regional varieties of UK English using the spoken part of the BNC, draws closer yet to the sociolinguistic approach, by taking into account one parameter of variation in speaker identity, namely region of residence. But large and (it may be surmised) relatively diverse groups of speakers are still grouped together under this approach.

So there is a difference in granularity between the Labovian study of speaker variation and the corpus-based study of text variation. However, it would be easy to exaggerate the significance of this difference. In many ways, variationist sociolinguistics and corpus linguistics are actually rather similar. Most notably, both take as their main data *actually observed language in use*. Like the early corpus linguists, variationist sociolinguists were doing this – working with transcripts of recorded speech – at a period in the history of linguistics when many researchers

were not. As McEnery and Wilson (2001: 13) observe, the Chomskyan rejection of empirical data was never adopted by phoneticians – including the sociolinguists we are discussing in this section; and Labov and those who followed in that tradition were working with real, observed data from the 1960s onwards (see, for instance, Labov 1972; Kerswill 1987, 1993; Tagliamonte 2007). Indeed, it is not uncommon for the dataset in a sociolinguistic study to be referred to as a 'corpus', although of course in absolute terms these datasets are much smaller than those constructed by even the early corpus linguists.

On methodological points, too, there has been considerable overlap between variationist sociolinguistics and corpus linguistics. Statistical methods, for instance, have been an integral part of the sociolinguistic toolbox for some time. For example, Cheshire (1982) uses tests of statistical significance on her data, and cluster analysis has been undertaken by a number of researchers (see, e.g., Le Page 1980; Le Page and Tabouret-Keller 1985; Kerswill and Williams 2000). The use of clustering in particular is clearly similar in spirit, if not in the mathematical nuts and bolts, to the statistical approaches applied in Biber's MD approach. In summary, then, we might say that the general approach to language and the methods applied are very similar in variationist sociolinguistics and corpus linguistics, but they differ in terms of the role taken in the analysis by phonetics, and in terms of the level (speaker versus text) at which variation is to be analysed. Is there, then, scope for interaction between these traditions of linguistics?

Clearly there is. On the one hand, more and more corpora are being constructed which pay attention to variation at the speaker level with a degree of detail that approaches that in a sociolinguistic analysis. For instance, the CANCODE corpus (see section 4.6) contains speaker metadata at this level. Smaller, more targeted corpora have also been constructed and used to examine speaker variation with respect to social aspects of language use. An example of this is Murphy (2009), who uses a 90,000-word corpus of spoken Irish English, balanced across age and gender, to analyse the pragmatics of the swearword *fuck*, thus building on the earlier, solely BNC-based work of McEnery *et al*. (2000a, 2000b) and McEnery and Xiao (2004b).

On the other hand, it is becoming more and more common for sociolinguistically sampled data to be compiled and described in a corpus-like manner. A good example of this tendency is the Newcastle Electronic Corpus of Tyneside English (NECTE: Allen *et al*. 2007; Beal *et al*. 2007). This corpus assembles data collected for two earlier sociolinguistic surveys, and incorporates phonetic transcriptions and audio recordings. But this data is encoded for distribution following standards established by corpus linguistics – using, for example, XML as the basis for markup, and with the transcripts being part-of-speech tagged and lemmatised. Similarly, analyses based on NECTE have been undertaken that utilise complex statistical methods extensively (Moisl and Jones 2005). Another example is the work done on the Linguistic Innovators Corpus by Gabrielatos *et al*. This corpus (Gabrielatos *et al*. 2010: 8)

> comprises 1.4 million words representing 100 hours of orthographically transcribed interviews with 118 speakers. The corpus was marked up for speaker turns (including backchannels and indications of speaker overlap), for the value of the extralinguistic variables built into the original samples . . . and for the friendship network score of each speaker . . . [12]

This study is of note for bringing an important concept from sociolinguistics to corpus linguistics – namely, friendship networks (see Milroy and Milroy 1992). Friendship networks are important because they allow researchers to explore what is obviously an important factor in language use – the influence of interlocutors, and members of peer groups in particular, on one another. Gabrielatos *et al*. find that these networks are a crucial explanatory factor. In their study of the prevalence of variation between *a* and *an* before words beginning in a vowel, they discover not only that the non-standard *a* + *vowel* usage is conditioned strongly by friendship networks, but also that it interacts with another item of metadata in the corpus, namely ethnicity, since 'the level of multiethnicity of the friendship networks that a speaker belongs to is a good predictor . . . of usage' (Gabrielatos *et al*. 2010: 24).

5.6 Summary

There has been a great deal of corpus-based work on variation across register, across time and across other 'external variables', and, as the four surveys we have undertaken in this chapter demonstrate, this work has been done from a number of contrasting perspectives. Much work on diachronic variation has been done using the general methodology of comparable corpus samples across time (using datasets such as the Brown Family, the Helsinki Corpus, or ARCHER). The study of synchronic variation, by contrast, differs substantially depending on whether social variables or register variables are the focus of the investigation. But in all cases the methods are typically both *quantitative* and *comparative* in focus. The single method that has been developed in the greatest detail is Biber's; while criticisms have been made of it, some of substance, these do not undermine the status of the MD approach as having contributed a great deal to our knowledge of language and especially the 'topography' of the speech–writing distinction.

Another notable feature of the work reviewed in this chapter is that it is all very firmly in what Tognini-Bonelli (2001) has characterised as the 'corpus-based' paradigm of research (see section 1.3), in that existing linguistic descriptions and/or theories are applied to datasets, and insights are gained from that data which develop or confirm the frameworks used to address the data. In the next chapter we will shift to look at a body of work that takes, by contrast, the so-called 'corpus-driven' perspective – namely, the work of the neo-Firthian school of corpus linguistics led by John Sinclair.

Further reading

Baker (2010) is essential further reading for this chapter as it reviews, in greater depth than is possible here, many of the issues touched upon in this chapter. Many good examples of corpus-based research on the history of English are collected in Rissanen *et al*. (1997) and Kytö *et al*. (1994). A comprehensive account of the diachronic research based on the Brown Family is provided by Leech *et al*. (2009). Rissanen (2008) is a useful discussion of the field as a whole.

For Biber's MD approach, the most detailed sources are Biber (1988, 1995a), or see Biber (1995b: 348) for a summary. However, Biber *et al*. (1998: 145–57, 278–80, *passim*) remains the most *accessible* introduction to the use of the MD approach; the more recent account by Biber and Conrad (2009) covers the MD approach but also many other aspects of variation across text types.

Practical activities

(A5-1) Choose three texts from contrasting genres – for instance, a chapter of a novel, a scientific research article, and a newspaper article. You can find texts on the web, or extract them from any corpus you have access to (the BNC, for instance, contains examples of all these genres). If possible choose part-of-speech-tagged texts, or else tag them yourself. Now, work out how to search for the following features using your concordancer – these are a subset of the sixty-seven features used by Biber (1988), including three features that are 'high' on each of three selected dimensions from Biber's analysis:

- Dimension 1 – involved (high) vs. informational (low)
  - 1st person pronouns
  - Present tense verbs
  - Contracted verb forms
- Dimension 2 – narrative (high) vs. non-narrative (low)
  - 3rd person pronouns other than *it*
  - Past tense verbs
  - Perfect aspect
- Dimension 5 – abstract (high) vs. non-abstract (low)
  - Passive voice
  - Past participial clauses
  - The adverbial subordinators *since*, *while* and *whereas*

Get frequencies (normalised per thousand or per million) for each feature in each text, and for each feature rank the texts in relative order. Does each feature on a given dimension produce the same relative order? (This will depend, above all, on precisely which texts you have chosen!) If so, does the relative order of your texts on that dimension seem functionally plausible? For instance, we would expect a novel to be more *involved*, more *narrative*, and less *abstract* than a scientific article.

(A5-2) Thinking about the task above: some of the features – especially passive voice and perfect aspect – are rather difficult to search for. How confident

are you that the search you devised returned *all* and *only* the examples of the structure you wanted? For example, for the perfect you need to search for any form of auxiliary *have*, followed by the past participle of any verb. However, it is possible for other words to come between these two components, or for the order to be inverted – compare *he has arrived* to *has he arrived?*, *he has not arrived* and *he has only just arrived*. Can you extend your search pattern to include examples like this? You may find that attempting to do this will also pick up other things that *aren't* instances of the perfect aspect – how much of a potential problem is this? Do similar concerns arise for the passive voice?

(A5-3) Get hold of a spoken corpus that represents at least two dialects of English. Choose a feature that is often said to differentiate the two dialects, search for it in the different dialect subcorpora, and see if the claimed distinction is actually in evidence in the frequency data from your corpus. Examples of things you might search for, depending on what dialects you have in your corpus, include the following (some easier than others, some impossible in a basic concordancer!):

- The use of dialect-specific vocabulary (some examples of such vocabulary may be found in Trudgill 1999: Chapter 5);
- Differing terms of address (e.g. in different regions of the UK, *mate*, *man*, *love*, *cock(er)* and many other terms are used as informal vocatives);
- The Northern English ditransitive (of the form *Joe gave it me* as opposed to Standard and Southern English *Joe gave me it*);
- Sequences of two modal verbs (often said to occur only in Scottish English);
- The present perfect (according to Sampson 2002, in the UK found frequently in Northern and Standard English, but less often in Southern English);
- The mandative subjunctive (*it is crucial that this be done*) in American versus British English;
- The *could care less / couldn't care less* idiom (by the usual account, both occur, and are synonymous, in US English; but only the latter occurs in UK English).

Questions for discussion

(Q5-1) One major problem that has been encountered when extending the Brown Family across time and across different languages is the stability of the genres in the sampling frame – even over relatively short periods and relatively similar cultures, genres can vary quite a bit. Given this issue, to what extent is it reasonable to expect long-period diachronic corpora to stick to the principles of *balance* and *representativeness* (see Chapter 1)?

What other practical problems might affect text sampling for a long-period diachronic corpus? What rules would you try to stick to if you were building such a corpus – and where might some leeway be needed?

(Q5-2) Biber's MD method, as reviewed in this chapter, is based largely on lexicogrammatical features – that is, the raw data consists of frequency counts of syntactic structures or (classes of) lexical items. Other levels of linguistic analysis are not generally present on the feature list – for instance, semantic, pragmatic or rhetorical features of a text. Can you think of any way such features could be added to the dataset for an MD analysis? Could any features of this type be included via associated lexicogrammatical features? If levels of analysis above the lexical, morphological and syntactic cannot be included in an MD analysis, do you think that this is a major problem for the MD methodology?

(Q5-3) Imagine that you are constructing a spoken corpus with a view to using it for a variationist-sociolinguistic analysis. What types of metadata would you wish to collect for each speaker in your corpus? What compromises might be necessary, for instance between the level of detail and the manageability of the final database? What ethical concerns would you need to bear in mind (you might want to refer also to Chapter 3 when considering this)?

6 Neo-Firthian corpus linguistics

6.1 Introduction

In this chapter, we will explore the approach to corpus linguistics taken by a group of scholars sometimes referred to collectively as *neo-Firthian*. As this label suggests, these researchers work within the framework of an approach to language suggested by J. R. Firth. The most prominent proponent of the neo-Firthian approach has been John Sinclair. Sinclair was one of the first people to bring Firth's ideas together with a corpus linguistic methodology (as Tognini-Bonelli 2001: 157 points out, Firth himself would probably not have subscribed to corpus methods); and Sinclair played a major role in enabling subsequent work along these lines. Many of the other key scholars in this tradition – including Michael Hoey, Susan Hunston, Bill Louw, Michael Stubbs, Wolfgang Teubert and Elena Tognini-Bonelli – are, or have previously been, associated with the University of Birmingham, where Sinclair was Professor of Modern English Language from 1965 to 2000.

Two central ideas in the approach to corpus linguistics favoured by neo-Firthians are *collocation* and *discourse*. It is, then, perhaps unfortunate that these terms are among the most multifariously defined – and, therefore, the most confusing – in all of linguistics. For this reason, in the next two sections we will examine some issues relating to the use of these terms, in theory and in practice. This will include discussion of how these terms are used both generally in linguistics and specifically in neo-Firthian corpus linguistics. Understanding the use of these terms is, to a large extent, key to understanding many of the positions taken by the neo-Firthians.

6.2 Collocation

Collocation is an old idea[1] and one that has been defined in various ways. In the twentieth century, the idea was brought into its modern form by Firth. In short, the term *collocation* denotes the idea that important aspects of the

meaning of a word (or another linguistic unit) are not contained within the word itself, considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which it frequently co-occurs, in what Firth (1968: 196) calls 'an abstraction at the syntagmatic level'.

However, as soon as we move beyond such basic generalities and attempt to pin down collocation either operationally or conceptually, we find a great multitude of different definitions. Let us firstly consider methodological or operational definitions. Most linguists today agree that the only way to reliably identify the collocates of a given word or phrase is to study patterns of co-occurrence in a text corpus. This contrasts with Firth himself, who 'only gave cliché examples [*sic*] like *You silly ass* or *He is an ass* as collocations of *ass*' (Esser 1999: 155), or early work by Greenbaum (1974), who used panels of native speakers to assess collocations by intuition. This use of intuition in identifying instances of collocation still persists – as Nesselhauf (2005: 282 fn65) notes, even modern corpus-based dictionaries such as *The Oxford Dictionary of Collocations* still rely heavily on the intuitions of lexicographers, used in concert with corpus data, to extract collocation information. For the purposes of this discussion, we will limit ourselves to definitions of collocation in terms of *co-occurrence patterns observed in corpus data*; thus, we will not consider any word co-occurrence phenomena identified by researcher intuition, by surveys of native speakers, or by means of any other type of data, to constitute *collocation* in the sense we are concerned with here.

Beyond considering the data used to identify a collocation, a further problem presents itself. Given the data, what does count, and what does not count, as a collocation? Different (groups of) practitioners, and different software tools, use the term *collocation* to refer to a wide range of different co-occurrence patterns that may be extracted from a corpus. For example, Harris (2006) uses the term to refer to recurring sequences of two or more words. These are more commonly referred to in computational linguistics as *n-grams* (i.e. sequences of *n* words) and in English Corpus Linguistics as *multi-word units*, *clusters* or *lexical bundles* (see section 4.5.2 for a discussion of lexical bundles); some examples of lexical bundles beginning in the word *cheese* are given in Table 6.1.

The other, perhaps more usual, family of approaches to collocation sees it as a potentially looser pattern of co-occurrence. In this case, a collocation is a co-occurrence pattern that exists between two items that frequently occur *in proximity* to one another – but not necessarily adjacently or, indeed, in any fixed order. Collocation in this sense may be considered a methodological elaboration on the concordance. Sinclair *et al.* (2004), a reprint of an early work by Sinclair from 1970, is a key text for the definition of collocation in this sense. Examining this study, we observe that Sinclair's basic methodological approach to collocation was fairly fixed at an early stage. Sinclair *et al.* (2004: 10) define *node* and

Table 6.1 *Three word n-grams ('lexical bundles') beginning with* cheese *with a frequency of ten or more in the written section of the BNC. Sequences involving punctuation have been excluded*

Three word n-gram	Frequency in written part of the BNC
cheese and a	23
cheese and wine	20
cheese and tomato	15
cheese with a	14
cheese on toast	14
cheese and onion	11
cheese and biscuits	10

collocate as follows:

> A node is an item whose total pattern of co-occurrence with other words is under examination; a collocate is any one of the items which appears with the node within a specified span. Essentially there is no difference in status between node and collocate; if word A is a node and word B one of its collocates, when word B is studied as a node, word A will become one of its collocates.

Collocates are also deemed to be determined within particular *spans* (Sinclair *et al*. 2004: 34):

> Two other terms . . . are span and span position. In order that these may be defined, imagine that there exists a text with types A and B contained in it. Now, treating A as the node, suppose B occurs as the next token after A somewhere in the text. Then we call B a collocate at span position +1. If it occurs as the next but one token after A, it is a collocate at span position +2, and so on.

Collocates are determined not simply on the basis of co-occurrence within a given span, though determining the maximum span to be examined for potential collocates is a significant filter to collocation. Rather, they are subject to a further filter which determines whether a collocation is significant or not (Sinclair *et al*. 2004: 35):

> The characteristics to be examined are the significant collocates of selected types. To explain what this means, imagine that we chose type A as a node, and a particular span position, say –*x*. We want to know whether a type B, say, is a significant collocate of A at span position –*x* . . .

The calculations of significance are tied to frequency data derived from the corpus (Sinclair *et al*. 2004: 28):

> The test of whether two words are significant collocates . . . requires 4 pieces of data; the length of the text in which the words appear, the number of times they both appear in the text, and the number of times they occur together.

It is with regard to this last point in Sinclair's view of collocation that we encounter a further area where definitions of collocations diverge. This approach places an emphasis upon both the role of frequency and the importance of a more-than-random co-occurrence in determining the significance of collocations. Significance, in this context, is clearly to be interpreted as *statistical* significance (that is, mathematical evidence that the co-occurrence pattern is unlikely to be due to chance: see section 2.6.2). Some neo-Firthian scholars – including Sinclair, for instance in his (2004b) study – do use statistics of this kind to calculate collocations, as will be discussed shortly; but especially in the earlier phase of the development of corpus linguistics, an alternative approach was frequently adopted. Sinclair in 1970 (Sinclair *et al*. 2004) introduces an impressionistic approach to identifying collocation, based on manually scanning through the concordance lines that result from a search for the node item. The later outline of collocation given by Sinclair (1991: 115–21) is fundamentally similar to the 1970 account. An example from Sinclair (2004a: 31) shows him using this technique:

> There are 154 instances of *naked eye* . . . By inspection of concordances, it is clear that there is a greater consistency of patterning to the left of the collocation than to the right, so we move our study step by step to the left. There is so much detail to be dealt with even in 151 lines that the main argument may get hopelessly obscured; hence this study is in two parts. The main argument is set out here with a few illustrative examples, and the discussion of the atypical, odd and wayward instances is returned to [subsequently] . . .

While frequency counts may be made (manually) of how often a particular pattern is observed relative to the number of concordance lines, in this approach statistical significance tests are generally not used. The use of such 'hand and eye' techniques in the analysis of corpus data is justified by Stubbs (1995: 27–8) on the following grounds:[2]

> Often, with quantitative linguistic data, no complex statistical procedures at all are necessary. It may be sufficient simply to count and list items. For example, in a corpus of 1.5 million words (LOB plus LUND), the following were noun collocates of *cause*, where f(n,c) is greater than or equal to 3:
>
> (17) accident, alarm, concern, confusion, damage, death, delay, fire, harm, trouble.
>
> It is obvious to the human analyst that these words are semantically related . . . Such raw frequencies require no further statistical manipulation to show a semantic pattern.

Elsewhere, Stubbs makes two other points against the use of statistical significance calculations in identifying collocations: firstly, that 'classical' statistical significance tests may make assumptions of randomness that are not true of language data (see also Kilgarriff 2005), and secondly, that in many cases where an analyst identifies a collocation, 'the levels of co-occurrence are so far above what one might expect by chance, that citing a probability level is rather pointless' (Stubbs 2001: 73–4).

We will here refer to the non-statistical technique as *collocation-via-concordance*. With this technique, it is the linguist's intuitive scanning of the concordance lines that yields up notable examples and patterns, not an algorithm or recoverable procedure. The computer's role ends with supplying the analyst with a set of (probably sorted) concordance lines. The linguist examines each line individually, identifying by eye the items and patterns which recur in proximity to the node word and reporting those that they find of note, possibly with manually compiled frequency counts but without statistical significance testing.

Collocation-via-concordance and the use of frequency data without significance testing are still common in neo-Firthian studies of concordance analysis. To give a concrete illustration of this, among twenty papers in a festschrift for Sinclair (Heffer and Sauntson 2000), eleven are substantively concerned with collocation;[3] but of those eleven, eight use no statistical tests, compared to two that report at least some statistical testing. By contrast, the final study of the eleven, Krishnamurthy (2000), strongly endorses the use of significance statistics and of analysis software that uses these statistics in the generation of collocation displays, arguing that this constitutes an improvement over the manual analysis which he characterises as 'obviously highly unsatisfactory'. Hunston (2002: 70) likewise supports the use of significance statistics, although she also, rightly in our view, cautions against an over-reliance on statistical evidence alone in determining the *meaning* of the results (Hunston 2002: 78–9). But Krishnamurthy and Hunston (and Sinclair 2004b: 45) appear to be generally in the minority among neo-Firthian analysts on this point. For instance, in a recent collection of papers by prominent neo-Firthians (Hoey *et al*. 2007), contributions by Hoey, Stubbs and Mahlberg all utilise the notion of collocation, though Stubbs and Mahlberg operationalise it in terms of n-grams, and all use frequency as evidence – but without significance testing.

Where work from a neo-Firthian perspective *does* identify collocations statistically, the intuition of the researcher is still regarded as the final arbiter of determining whether or not a specific candidate collocate is indeed a collocate. For instance, Stubbs (2001: 66–7) distinguishes between first order data (raw corpus data), second order data (corpus data as manipulated by a basic concordance program) and third order data (corpus data that has been manipulated using statistical analyses to present patterns within the data). Stubbs argues that the human analyst must have primacy over third order data, constantly checking it against first and second order data – that is, checking the outcomes of statistical calculation of collocation against concordances and the raw text (Stubbs

2001: 71). Otherwise, co-occurrence patterns realised by a range of different, individually infrequent co-occurring wordforms may be missed, whereas using the collocation-via-concordance method they are easily identified. So, in sum, while the utility of statistics for the extraction of collocation is acknowledged, and indeed central to Sinclair *et al*.'s (2004) early account, in practice for most neo-Firthian analysts it remains subordinate to the linguist's intuitions and hand-and-eye methods. For other schools of corpus linguistics, reliance on statistical testing to identify collocations is much more prevalent. This is partly due to pragmatic considerations – with large datasets, calculating and manipulating co-occurrence frequencies automatically saves time and extends the scale of analysis that is feasible – but this approach also allows analysts to be much more explicit about the criteria used to determine whether or not a specific word is a collocate of a given node.

If we accept that statistical tests applied to frequency data should have the key role in determining collocation – 'collocation-via-significance', as we might dub it – still further issues arise. One relates to the choice of statistic. The most usual way to automate collocation-via-significance is by comparing the frequency of each word within the window of text defined by the span around the node word, against its frequency in the rest of the corpus. If the difference between the frequencies is sufficiently great, the word being examined is identified as a collocate of the node word. Often, this is done by using a significance test to see whether there is a sufficiently low probability that the difference between the frequency-with-node and the frequency-elsewhere is a purely random effect (see also our preliminary discussion in section 2.6.2). The tests used may be precisely the same as those used in statistical significance testing generally – chi-squared, log-likelihood and so on. But other statistics can be used as well (such as the t-score, z-score or mutual information); reviews of the statistics and the differences among them may be found in Hunston (2002: 70–5), Hoffmann *et al*. (2008: 149–58) and Baker (2006: 101–3). The variety of measures which may be used to determine the significance of a collocation is clearly problematic, for, as Table 6.2 demonstrates, the list of collocates returned is determined in part by the statistic used to calculate them.

There is very little similarity between the two lists in Table 6.2; the rankings are completely different, and only two words, *cheddar* and *parmesan*, appear in both lists – although we might find more overlap if we considered longer collocate lists, of course. Thus, calculating collocation via statistical testing, uncontroversial in theory, becomes problematic in practice. Because the analyst's choice of statistic has such a major effect on the outcome, there is in effect an inherent subjectivity in the determination of what is, and what is not, a collocate. Other choices by the analyst can also alter the results. For example, Table 6.3 shows how the list of collocates for *cheese* in the written BNC alters when we change the span from which collocates are extracted from three (i.e. three words before or after) to five.

Again, the ordering is different and only two collocates are shared in common. Additionally, the prevalence of grammatical words in the top ten collocates when

Table 6.2 *The top ten collocates of* cheese *in the BNC calculated using different statistical measures (intra-sentential collocates within a span of +/–3 only)*

Top ten collocates of *cheese* (in descending order) in the written BNC using the log-likelihood measure	Top ten collocates of *cheese* (in descending order) in the written BNC using the mutual information measure
bread	*feta*
cream	*ricotta*
and	*parmesan*
grated	*mozzarella*
cottage	*cheddar*
butter	*gruyère*
milk	*macaroni*
cheddar	*unpasteurised*
parmesan	*fondue*
wine	*appenzell*

Table 6.3 *The top ten collocates of* cheese *in the BNC calculated using the same statistical measure (log-likelihood) but different spans*

Top ten collocates of *cheese* (in descending order) in the written BNC, with a span of 3	Top ten collocates of *cheese* (in descending order) in the written BNC, with a span of 5
bread	*oz*[*]
cream	*milk*
and	*cheese*
grated	*bread*
cottage	*lunch*
butter	*a*
milk	*or*
cheddar	*as*
parmesan	*with*
wine	*like*

[*] A note for puzzled readers – *oz* is an abbreviation of the imperial measurement *ounce* (0.028kg).

the span is five suggests that this wider span produces a different *type* of collocate than that yielded by a span of three.

The problems inherent in determining collocations are, of course, known. The issue of collocation statistics has been visited and revisited many times in the past few decades, often in the broader context of work in computational linguistics looking at such issues as term extraction and word similarity

measurement – issues which, like collocation, require statistical testing (see Church and Hanks 1989; Dunning 1993; Stubbs 1995; Daille *et al*. 1996; Lin 1998; Mason 1999; Curran 2004; Evert 2005; Petrovic *et al*. 2006). Similarly, the use of specific spans for collocation has varied over time with authors using spans of two (Clear 1993), three (Gledhill 2000), four (Sinclair *et al*. 2004: 13) and five (Huang *et al*. 1994; Xu *et al*. 2003; Stuart and Trelis 2006) words, though in fairness the majority of corpus linguists working on English have adopted Sinclair's guideline of a span of +/– 4. As Stubbs (2001: 29) points out, '[t]here is some consensus, but no total agreement, that significant collocates are usually found with a span of 4:4'. It should, however, be noted that in computational linguistics there has been a 'common practice of using a five-word span for collocate searching' (Seretan and Wehrli 2007: 75); and when collocation is operationalised in terms of n-grams, larger spans may be used (e.g. Mahlberg 2007c analyses eight-word clusters). However, it has also been suggested that collocation should not be controlled by fixed-length word spans, but that it should, rather, be calculated with regard to the syntactic structures within which the node word appears. The possibility of such an approach was noticed very early in the history of the corpus-aided study of collocation, by Berry-Rogghe (1973), who, however, rejects it as undesirable. More recently, arguments have been made in favour of this approach. For example, Grefenstette (1992: 90) notes that:

> [u]se of syntactic analysis opens up a much wider range of contexts than . . . co-occurrence within a window . . . Syntactic analysis allows us to know what words modify other words, and to develop contexts from this information . . .

Acknowledging and annotating those syntactic structures prior to the search for collocates, and actively using that annotation as a control on the identification of collocates, is something that linguists and tools developers are actively exploring.[4] This syntactically informed approach to collocation is extended by Stefanowitsch and Gries (2003) to a novel theoretical construct dubbed *collostruction*, the mutual attraction or repulsion between slots in syntactic structures and lexical items; this will be reviewed in depth in the next chapter. For now, the main point is that identifying collocates by proximity, within a short span of words to the left and right of the node, may simply be a means of gaining an imperfect view of how the meaning of the node relates to the words (and syntactic structures) around it.

The variability in the process of determining what a collocation is should not be viewed as a marginal criticism – it has been the subject of intense debate over a prolonged period of time. It certainly should lead us to agree with Esser (1999: 155), who notes that '[a]lthough the notion of collocation is frequently used it often remains mysterious' and that it is 'difficult to operationalise'. We have only explored one or two areas where the procedure of identifying a collocation may vary. Others exist, including: whether or not collocations should be identified across sentence boundaries; whether collocation should be calculated on the basis of wordforms, lemmas or multi-word units; what the minimum frequency cut-off should be below which a co-occurring word will not be considered as a

potential collocate; and whether collocates should be explored within one or many genres.[5] This procedural variability has a significant consequence for studies of collocation – namely, comparing the results of different studies of the same node word can be difficult, and may indeed be impossible, as the basis on which the collocates of that node were determined may vary quite significantly.

We have seen, then, that there are many variations in the definition and operationalisation of collocation, the most notable being the distinction between collocation-via-concordance and collocation-via-significance, and that there is a tendency among the neo-Firthian school of corpus linguists to favour the former of these. We will argue, later in this chapter, that the preference of Sinclair and his colleagues for collocation-via-concordance over automated statistical collocation is not incidental, but is, rather, crucially linked to their theoretical stances on the nature of language and of corpus linguistics. In particular, a number of extensions to the idea of collocation which are typically identified using the collocation-via-concordance approach – namely colligation, semantic preference and semantic prosody – have become central concepts in neo-Firthian corpus linguistics. We will return later to semantic preference and semantic prosody, and comment here only briefly on colligation.

It may often be observed that a word collocates not only with some meaningful lexical items, but also, or even instead, with some grammatical markers or grammatical categories. Such collocations are referred to as *colligations*. For instance, many words colligate with the word *the*, which is a grammatical marker of definiteness rather than a word that carries significant semantic content. Colligation with *the* is in fact typical of many words traditionally described as nouns. Likewise, many of the individual words that collocate with the node word *made* turn out to be auxiliary verbs (such as *had*, *have*, *been* and so on).[6] So we can say that *made* colligates with the grammatical category of auxiliaries. But colligation is not simply a matter of co-occurrence with particular parts of speech. Patterns of consistent co-occurrence of a word with different syntactic contexts are also described as colligation; for instance, Hoey (2005: 48–9) identifies the preference of the noun *consequence* to be head of a noun group, rather than part of the pre- or postmodification of the head, as a colligational property. As with collocation, contrasting ways of accessing colligation have developed. In keeping with the practices described above, at times colligates are identified without recourse to statistical tests (see, for example, Hoey 2005: 77–8), though they may also be identified by looking for grammatical words on a statistically generated list of collocates. But yet other approaches exist. For example, in the SARA concordancing software and the Xaira system derived from it, both originally created for searching the BNC (see section 2.5.3), 'colligates' are defined as part-of-speech tags that are particularly common in the vicinity of the node (as opposed to collocates, which are *words* that are particularly common in the vicinity of the node). In Table 6.4 we present some data on the colligates of *cheese* using the two different methods. Some of the results overlap (e.g. *and* as the top grammatical word, and CJC – conjunction – as the top tag). Others do not (e.g. the link to NN0, which generally indicates measurements or amounts).

Table 6.4 *Top ten colligates of* cheese *in the written BNC calculated using a word-based and a tag-based approach (statistic: log-likelihood; span: +/− 3, intra-sentential collocates only)*

Top ten collocates of *cheese* in the sense of 'grammatical words that collocate' (in descending order)	Top ten colligates of *cheese* in the sense of 'co-occurring part-of-speech tags' (in descending order)*
and	CJC
with	PUN
or	NN1-AJ0
a	NN0
some	VVB-NN1
&	NN1
until	VVB
into	AJ0-VVD
is	VVD-AJ0
over	VVN-VVD

* The tags are from the CLAWS5 tagset used to tag the BNC: http://ucrel.lancs.ac.uk/claws5tags.html.

As neo-Firthian linguistics is generally opposed to corpus annotation in the first place, as we will discuss later in this chapter, the tag-based method is not directly compatible with the type of colligational analysis undertaken by Hoey and others, in spite of the fact that both are dubbed *colligation*.

So far, we have discussed methodological definitions of collocation (and colligation) – that is, the different ways in which the idea that Firth proposed is operationalised in contemporary corpus linguistics. However, there is another aspect to the definition of this phenomenon, which is the ontological status of collocations. In other words, what exactly is their nature? Where – and how – do they exist?

Some theories of linguistics have not assigned collocation any notable status at all. An example of such a theory would be the generative (Chomskyan) approach, where words are inserted individually into the slots in a syntactic structure generated by formal, algebraic rules, which themselves make no reference to meaning. However, neo-Firthians argue – and the overwhelming majority of contemporary corpus linguists would agree – that a word's collocational patterns are a crucial part of its meaning. There are differences of opinion, however, on the precise nature – and the degree of centrality – that collocation has, or should have, in the theory of language. Although neo-Firthians typically lay more emphasis on collocation than others, there are major conceptual differences even among neo-Firthians on this point.

In Sinclair's writings, the centrality of collocation is linked above all to the centrality of meaning. Like the notion of collocation itself, the assignment of a central role in linguistics to *meaning in context* ultimately derives from Firth but

was developed extensively by Sinclair. Briefly, Sinclair (2004a: 18–20) argues from the prevalence of collocation in text that meanings in running text are not confined to individual words but have wider extents – and the beginning and end points of the expression of a particular meaning may not be evident. From the prevalence of collocation also follows the notion that words are not individually selected by speakers. Rather units of meaning are selected, and each unit of meaning brings along several words. In sum, for Sinclair, 'a grammar is a grammar of meanings and not of words' (2004a: 18) and '[t]he meaning of words together is different from their independent meanings' (2004a: 20). The units Sinclair argues for, units which reach beyond the word and thus incorporate the collocations of words, are referred to either as *extended units of meaning* or as *lexical items* (Stubbs 2001: 60). Stubbs (2001: 87–9) develops Sinclair's ideas into a systematic account of how the extended lexical units around a word may be studied by the successive analysis of collocations, colligations, semantic preferences and discourse prosodies (also called *semantic prosodies*: see below). Colligation, semantic preference and discourse prosodies are all abstractions of collocation – that is, they are built upon a collocation analysis. So in the form of neo-Firthian analysis proposed by Sinclair, collocation becomes the absolute keystone of the study of lexis and semantics, and thus (due to the centrality of meaning in Firthian linguistic theory) the keystone of the entire approach to language.

Nor is this the only form of neo-Firthian theory in which collocation is central. Another example is the theory of Lexical Priming (Hoey 2005, 2007a, 2007b), which adopts a different perspective on the phenomenon of extended lexical units. Hoey proposes that collocation in texts (including spoken language) is both the cause of, and the result of, an underlying phenomenon in the mind. This phenomenon is dubbed *lexical priming* by Hoey. Words are *primed* to co-occur with other words, so that a person perceiving word A is then psychologically primed (i.e. predisposed in some way) to anticipate one or more of the words that A is linked to in the mind; likewise a person producing the word A is psychologically primed to subsequently produce those words.

We will not go further into the details of Hoey's theory at this juncture, although we will return to it in section 6.5.3. But even from this cursory overview, it should be clear that under Hoey's theory of the language system, collocation is at root a mind-internal phenomenon, albeit one that must be discovered in the discourse via the methods of corpus linguistics. However, Hoey's theory is at odds with the views of some other neo-Firthians. They have adopted a radically different stance, best summed up by the following declaration by Teubert (2005: 2–3):

> The focus of corpus linguistics is on meaning. Meaning is what is being verbally communicated between members of a discourse community. Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language. It claims no privileged knowledge of the workings of the mind or of an innate language faculty.

It would seem, then, that the ongoing evolution of approaches to collocation has led to the emergence of at least two distinct schools of thought in neo-Firthian linguistics. Setting aside the methodological issues, the key fault line is the nature of the collocation. For those who would identify with Hoey, the collocation is a window into a mind-internal phenomenon. As will be seen in Chapter 8, Hoey's focus brings neo-Firthian corpus linguistics into closer contact and potential rapprochement with mainstream theoretical linguistics and psycholinguistics. For those who would identify with Teubert, corpus linguistics has nothing to contribute to work on mind-internal phenomena and instead is a tool for exploring discourse. Considering the central role that discourse has to play in this version of corpus linguistics, the next section will explore the neo-Firthian approach to discourse in some detail.

6.3 Discourse

Having discussed at length the varied definitions, conceptualisations and operationalisations of *collocation*, let us now turn to *discourse*. The only consistent understanding of this term across all fields of linguistics is that it refers to units of language above the level of the sentence. However, there exists a number of different ideas about the nature of discourse, and the approaches that should be taken to its investigation. In functional-typological grammar, the analysis of *discourse* is undertaken to understand aspects of grammar that are motivated by patterns that stretch over many sentences. For example, it has been proposed by Du Bois (1987) that the grammatical phenomenon of *ergativity* has a 'discourse basis' (see section 7.4). We will return to functional grammar, and its interaction with methodologies of corpus linguistics, in the next chapter. Here, we are only concerned with comparing functional grammar's fairly low-level definition of *discourse*, with which the reader is probably familiar, with two contrasting, less restricted definitions of the term, used in neo-Firthian corpus linguistics and in the discipline of Critical Discourse Analysis.

Critical Discourse Analysis (CDA) studies *discourse* in a very broad sense. A *discourse* in this sense is not merely a group of sentences, a text or a class of texts. It is also a *practice*: a characteristic type of language use found in a group of texts, or at large in the language of a community. Further, *discourses* are not only ways of talking about something, they are also ways of *thinking* about it. By studying language, critical discourse analysts aim to uncover how groups of people conceptualise themselves, their social setting, other groups of people and the issues that matter to them. Thus, CDA is an overtly sociological and political enterprise, with an acknowledged political (and typically left-wing) stance, as evidenced by a frequent concern with power relations among groups within society and the treatment, in particular, of relatively less powerful groups of people. The critical discourse analyst typically approaches their goal via close

analysis of a text or texts, but other techniques for the investigation of discourses in the broader sense are also used, for example focus groups (see Cahnmann *et al*. 2005). Particularly in recent years, some CDA-oriented studies have been undertaken utilising corpus resources and corpus methods (see section 1.6.2).

If this is *discourse* in the context of CDA, what is *discourse* in neo-Firthian corpus linguistics? One immediate difference is that Sinclair's approach to discourse, unlike the CDA approach to discourse, claims to be expressly apolitical – he argues that even emotive issues (the example he analyses is press coverage of a legal case involving the unavoidably lethal separation of a pair of conjoined twins; Sinclair 2004a: 120–6) should be approached as if the analyst does not have an opinion. This is in very clear contrast to the CDA approach, where the political stance of the analyst is overt and often explicit. All in all, however, Sinclair's use of the term *discourse* can probably best be characterised as in some way intermediate between the traditional meaning of the term found in functional linguistics (language analysed across sentence boundaries), and the broader meanings of the term found in CDA (ways of talking and thinking about something). It is, to be certain, more traditional than the usage adopted by CDA practitioners. This is not surprising, as Sinclair's interest and involvement in the analysis of discourse began in the early 1970s, whereas CDA as a distinct school of linguistics did not emerge until the mid-to-late 1980s.

Sinclair looks at discourse in the particular sense of the *structure* of a text, and in terms of issues such as cohesion and coherence. The model he develops (see, e.g., Sinclair 2004a: 82–101) is particularly concerned with how each sentence in a text relates to those that precede and those that follow. To capture a sentence's relationships with the preceding text, he argues for the concept of *encapsulation*, defined as follows (Sinclair 2004a: 83–4):

> each new sentence refers to the previous one by an act of reference. By referring to the whole of the previous sentence, a new sentence uses it as part of the subject matter. This removes its discourse function, leaving only the meaning which it has created . . . The current sentence would encapsulate the previous one, which in its turn had encapsulated its predecessor, and so on back to the beginning of the text . . . Any sentence, then, would be a precise manifestation of the whole text up to that point.

As well as encapsulating, sentences can 'prospect', linking forwards in the text: '[p]rospection occurs where the phrasing of a sentence leads the addressee to expect something specific in the next sentence' and in this case, '[t]he sentence fulfilling the prospection does not encapsulate the prospecting sentence' (Sinclair 2004a: 88, 97). The links forwards and backwards between sentences make up the structure of the text. Sinclair's analysis of discourse in this sense is done at the level of the individual text, rather than using corpus data, although Sinclair saw his work on discourse and on corpus linguistics as closely linked (Sinclair 2004a: 10).

Given Sinclair's long-standing concern with the nature of discourse (in this particular sense), it is perhaps not surprising that the study of discourse has been

and remains a primary concern of the neo-Firthian school of corpus linguistics. This does not necessarily mean taking precisely Sinclair's approach. For example, Teubert defines discourse rather more broadly as 'the entirety of all the utterances of a discourse community', where the discourse community includes 'all those who have contributed and are contributing, through their utterances, to the global discourse, ever since mankind began contributing utterances to the discourse'; by contrast, '[a] discourse that can be made the object of a linguistic (or sociological) investigation is a construct' (Teubert 2007a: 73, 76). Meanwhile, for Stubbs (2007: 145) discourse is defined as 'intentional and meaningful social action, which cannot be reduced to physical behaviour or its traces in text'.

The conceptualisation of *discourse* in neo-Firthian linguistics, although varied, is thus clearly very different from both the other accounts of discourse that were briefly expounded above – the view taken of discourse in functional-typological grammar, and the view taken in CDA. In particular, whereas for CDA practitioners *discourses* are (among other things) ways in which people and groups of people think about some aspect of the world, some neo-Firthian corpus linguists, notably Teubert as discussed already, have explicitly denied the view that the study of discourse can tell us anything about mental concepts or how people think. But having pointed out the contrasts between CDA and neo-Firthian discourse analysis, it is worth pointing out a feature that, in fact, they share. This is an influence from the poststructuralist school of literary and cultural criticism (Teubert 2007a: 73 links his view on discourse to that of Foucault, for instance). Poststructuralists consider *discourse* (again, using the term in a slightly different sense) to be the central, and to some degree the only possible, object of study: the discourse is not studied to find out about 'the real world', but rather to find out how 'the real world' is talked about. This is a feature of much CDA, and also of some neo-Firthian approaches; in some cases, a connection between the discourse and a reality external to the discourse is actually denied, as exemplified by Teubert's (2005: 3) comment:

> Meaning is in the discourse. Once we ask what a text segment means, we will find the answer only in the discourse, in past text segments which help to interpret this segment, or in new contributions which respond to our question. Meaning does not concern the world outside the discourse. There is no direct link between the discourse and the 'real world'. It is up to each individual to connect the text segment to their first-person experiences, i.e. to some discourse-external ideation or to the 'real world'. How such a connection works is outside the realm of the corpus linguist.

6.4 Semantic prosody and semantic preference

The concept of semantic prosody was originally outlined by Louw (1993). Others have produced expansions, and variations, on Louw's original characterisation of this phenomenon; furthermore, as we will discuss at the

end of this section, the concept has been subject to considerable criticism and debate.

Semantic prosody – also referred to as *discourse prosody* by authors following Stubbs' (2001) usage – is a concept rooted in the neo-Firthian concordance-based analysis of collocation. It may be understood as a concept related to that of *connotation* in more traditional approaches to semantics. Words or phrases are said to have a negative or positive semantic prosody if they typically co-occur with units that have a negative or positive meaning. So for instance, the lemma *happen* occurs more frequently than would be expected by chance alone with subjects that can be evaluated as negative. In twenty randomly selected concordance lines which we took from the BNC, for example, something evaluated negatively *happened* in eight cases, something evaluated positively *happened* in four cases, and the other eight were either neutrally evaluated or were ambiguous. The negative things are not, necessarily, themselves significant collocates of *happen*; it is when they are considered in the aggregate that their frequency becomes notable. Thus, an analysis of semantic prosody is an abstraction across multiple, different contexts of usage.

The key difference from the traditional notion of connotation is that the semantic prosodies are not necessarily accessible to intuition, which is often used to make judgements about the connotations of a word. Rather, a semantic prosody can *only* be discovered by analysis of a concordance, as Louw (1993: 159) argues. While semantic prosodies may in some cases explain connotations which we have previously sensed intuitively to exist, in other cases the positive or negative associations of a given lexical item may not be accessible from our conscious knowledge (Hunston 2002: 142). Indeed, Tognini-Bonelli (2001: 114) argues that 'semantic prosodies are *mainly* engaged at the subconscious level' (our emphasis).

The metaphor present in the term *semantic prosody* – namely, characterising the domain of lexis and semantics in terms of the domain of intonation and nonsegmental phonetics – is no coincidence. Rather, it is an expression of the continuity of neo-Firthian corpus linguistics with Firthian phonology. Rather than focusing on individual phonetic segments in terms of phonemes and allophones, as the then-dominant American Structuralist approach to phonology usually did, Firth placed a significant emphasis on how sounds work in context to create meanings. He used the term *prosody* for the many ways in which a sound may be influenced by its environment. For instance, while in English nasalisation is a consonantal feature that distinguishes only the three stops [n], [m] and [ŋ], Firth noted the prosodic phenomenon whereby *vowel* segments in the neighbourhood of one or more of these stops become nasalised (e.g. the vowel in *man* is normally pronounced as [ã] not [a]). The notion of semantic prosody is intended to be directly parallel to this, as outlined in Louw's (1993) paper, which introduced the idea (although Louw credits the actual concept to Sinclair; Stewart 2009, however, argues that Louw's and Sinclair's treatments of the topic are very different). Louw defines semantic prosody as '[a] consistent aura of meaning with which a form is imbued by its collocates' (1993: 157), and argues that the

habitual collocates of a form are 'capable of colouring it, so it can no longer be seen in isolation from its semantic prosody' (1993: 159). Louw invokes semantic prosody as a means to explain certain stylistic effects, in particular *irony*, which in this context may be defined as the creation of an impression on the part of the reader that the author does not agree with the apparent import of their own words. For example, Louw considers a passage from a novel by David Lodge where certain people are described as 'austerely bent on self-improvement'. Noting that a concordance of *bent on* shows that the things that people are 'bent on' are typically negatively evaluated, Louw argues that a collocational clash between the expression's usual semantic prosody (negative) and the typical evaluation of *self-improvement* (positive) creates the effect of irony. By using the term *bent on*, Lodge conveys the message that *self-improvement* is to be considered a bad thing in this context.

What kinds of terms have been identified as carrying notable semantic prosodies? The earliest example in the literature may be the negative semantic prosody of *happen*, discussed above, which was noted as early as the mid-1980s by Sinclair (1987). Other terms identified as having negative semantic prosodies include *set in*, *utterly* (Louw 1993), *cause* (Stubbs 1995), *undergo* (Stubbs 2001), *occur*, *come about*, *take place* (Partington 2004) and *persistent* (Hunston 2007). Terms with positive semantic prosodies are more rarely discussed; one example is *provide* (Stubbs 1995). There is a relative dearth of research into semantic prosodies of words in languages other than English. Tognini-Bonelli (2001: 113–16) presents evidence that the Italian verb *andare incontro*, 'move towards', has a negative semantic prosody. Likewise Xiao and McEnery (2006) demonstrate that the phenomenon can also be observed in Mandarin Chinese – and in many cases that there is cross-linguistic comparability in the prosodies; so, for instance, several Mandarin words that are possible translations of English *cause* (such as *zao4cheng2*) also have the negative semantic prosody characteristic of English *cause*.

One long-standing problem with the notion of semantic prosody is discriminating it from the phenomenon of *semantic preference*. Stubbs (2001: 65–6) defines semantic preference as 'the relation, not between individual words, but between a lemma or word-form and a set of semantically related words' and notes that there is a fuzzy boundary between the two concepts. To put it another way, just as colligation (discussed above) groups elements of the text around a node item on the basis of grammatical similarity, semantic preference groups elements on the basis of semantic similarity. Semantic preference and semantic prosody are similar in that both are abstractions across multiple, different collocations. And indeed, some work in the field has confused them (Hoey 2005: 23 notes that some phenomena which he identified in earlier work, e.g. Hoey 1997, as semantic prosodies are actually better classed as semantic preferences, or *semantic associations* in his terminology). How then are they to be distinguished? There are, in general, two distinctions that are typically made. One is that, whereas a semantic preference may be in favour of any definable semantic field, a semantic prosody is always either for positive evaluation or for negative evaluation. The

instances of semantic prosody discussed by Louw (1993), for example, are of this type. Partington (2004: 150) attributes this distinction to Sinclair, but in fact this attribution is problematic. In the very paper that Partington cites, Sinclair (1999: 33–4) does indeed define semantic prosody as 'attitudinal' but also argues for semantic prosodies that are more specific than merely positive or negative evaluation. For instance, he argues that the expression *the naked eye* has a semantic prosody of 'difficulty'. This is clearly more semantically specific than mere negative evaluation. Perhaps more crucial to the distinction between semantic preference and semantic prosody is Sinclair's (1999: 34) observation that semantic prosody is 'on the pragmatic side of the semantics/pragmatics continuum. It is thus capable of a wide range of realisation.'

In other words, semantic preference links the node to some word in its context drawn from a particular semantic field, whereas semantic prosody links the node to some expression of attitude or evaluation which may not be a single word, but may be given in the wider context. Sinclair's analysis of *the naked eye* makes the distinction very clear. From concordance lines such as *too faint to be seen with the naked eye*, *it is not really visible to the naked eye* and *cannot always be perceived by the naked eye*, Sinclair argues that *the naked eye* has a preference for the semantic field of visibility, evident in words such as *seen*, *visible* and *perceived*. The semantic prosody of *difficulty*, by contrast, is not evident from single words in the context of the node, but rather by a pragmatic interpretation (by the reader or by an analyst) of extended sections of co-text such as *too faint to be seen*, *it is not really visible* and *cannot always be perceived* – where there are no particular individual words that can be identified as belonging to the semantic field of *difficulty*. On this basis, as Hunston (2007: 258) points out, it is clear that Sinclair sees semantic prosody as 'a discourse function of a sequence rather than a property of a word'. This is in contrast to some others who have employed the concept (e.g. Gabrielatos and Baker 2008 use an approach to semantic prosody based solely on abstractions across multiple individual-word collocates). The importance of discourse and pragmatics to the concept of semantic prosody may lead us to agree with Stubbs (2001: 65–6), who proposes the alternative term *discourse prosody* for the same phenomenon, on the basis that it is concerned with speaker meaning (pragmatics) rather than word meaning (semantics – note that it is clearly with word meaning that semantic *preference* is concerned). Possibly 'pragmatic prosody' would be an even more appropriate label. However, in spite of Stubbs' persuasive argument, *semantic prosody* is the term most widely used in the literature, and thus we will continue to use it here.

Semantic prosody, and in particular Louw's account of it, has been refined, questioned and criticised in various ways. Hunston (2007), for instance, provides substantial evidence that semantic prosodies may be bound to particular registers; that the semantic prosodies of a particular item are associated with particular phraseologies, and if the phraseology around an item changes, the semantic prosody may change too; and that the presence of a semantic prosody may

be dependent on the presence of some other grammatical or semantic feature alongside the node (for instance, *persistent* has a negative semantic prosody only as an attributive adjective, and *cause* lacks its normal negative semantic prosody in contexts that do not refer to human beings). Hunston's findings imply that a degree of caution is required in making and evaluating claims that a particular word or phrase 'possesses' a particular semantic prosody. A set of more serious criticisms have been aimed at the concept of semantic prosody by Whitsitt (2005). Some of Whitsitt's criticisms are ill-founded – for instance, in our judgement his attacks on the analogies and metaphors that Louw (1993) employs in outlining the nature of semantic prosody, whether or not they are accurate in substance, do not amount to an invalidation of the concept, which must stand or fall on its own merits rather than those of the analogies used to present it.

However, Whitsitt does raise some important concerns. In particular, he focuses on the diachronic implications of Louw's view of semantic prosody. The issue here is one of the causation of semantic prosodies. Louw presents the semantic prosody of an item as being *caused* by its normal contexts. That is, because a word normally occurs in (for instance) contexts with negative meaning, it becomes imbued with an element of that negative meaning and can then carry that negative meaning over to other contexts. The logical implication of this causative view, made implicitly by Louw's description of it as a process of 'imbuing', is that usage in negative contexts has caused over time a change in the meaning of the word. It has gone from initially having a simple, pragmatically neutral meaning to subsequently having, in addition, a semantic prosody. Whitsitt accurately points out that concordances from a corpus of contemporary English cannot, by their very nature, provide evidence for such a process of change over time: 'no matter how long one looks at what is the synchronic use of a word like *set in*, there is no evidence for assuming that we can see the results of a diachronic process of *imbuing*' (Whitsitt 2005: 296; see also Stewart 2009: Chapter 3). From the same synchronic evidence, we could argue equally well that the chain of causation moves in the opposite direction: that a word possesses its semantic prosody *ab initio*, and its appearance in typically positive or typically negative contexts is a *result of* rather than the *cause of* its semantic prosody. Hunston (2007: 266) concurs partially with Whitsitt on this point, noting that if 'meaning does not exist except in context', then it is illogical to say 'that a word or phrase can carry its meaning across from one text to another' but that, '[o]n the other hand, while meaning may not be transferable from one text to another, resonances of intertextuality are difficult to deny'. She resolves this dilemma by noting a subtle distinction between the *explanatory* and *predictive* uses of semantic prosody. As a way of accounting for a reader's reaction to some aspect of a text, the 'resonances of intertextuality' characterised as semantic prosody are a powerful explanatory principle. However, if the claim that a word has a given semantic prosody is taken as a prediction that the meaning of that semantic prosody must *always* be present when that word is used – which is what Louw's (1993) claim that semantic prosody can be used to identify irony necessarily implies – it becomes

problematic. As Hunston notes, there can always be found counterexamples to the semantic prosody of any given item, i.e. instances where it occurs in a context that lacks the typical negative or positive meaning.

It is in what Hunston terms the 'predictive' sense that the most grandiose claims have been made for semantic prosody. These claims are most clearly associated with Louw. In particular, Louw originated the notion that '[semantic prosodies] operate in a binary relationship between irony and insincerity' (Louw 2000: 53). This is explained in the extended working definition of semantic prosody that he offers (Louw 2000: 57):

> A semantic prosody refers to a form of meaning which is established through the proximity of a consistent series of collocates, often characterisable as positive or negative, and whose primary function is the expression of the attitude of its speaker or writer towards some pragmatic situation. A secondary, although no less important attitudinal function of semantic prosodies is the creation of irony through the deliberate injection of a form which clashes with the prosody's consistent series of collocates. Where such reversals are inadvertent they are indicative of the speaker's or writer's insincerity.

That is, Louw's 'binarity' claim is that in any context where a form that has a negative semantic prosody (for instance) co-occurs with a positive form (for instance), rather than its typical, negative collocates, this is indicative of falsehood on the part of the writer – either overt falsehood (irony) or covert falsehood (insincerity). He illustrates this with an example drawn from Moon (1998: 161, 256–7): *President Clinton fanned the flames of optimism in Northern Ireland*. The normal semantic prosody of *fan the flames of* is negative. Louw (2000: 53) argues that:

> [t]he conclusion we reach as we unravel this line becomes an act of critical literacy. For a split second the form is rejected as incoherent on the basis that *optimism* bears no resemblance to the normal collocates of this fixed expression. However, within a further split second, the critical message of the writer is unravelled: the peace process is, *ironically*, almost as aggressive as the war it is designed to end. The line has to be an intentional comment on US foreign policy.

The intertextuality noted by Louw is certainly present. However, what is questionable is the claim that the implication about the Northern Ireland peace process which he perceives is *necessarily* an 'intentional' comment, i.e. necessarily revelatory of the author's true feelings. It is trivially easy to demonstrate that the reproduction of a fixed linguistic pattern whose meaning runs contrary to the speaker's own expressed opinion is not evidence that that opinion is expressed insincerely. For instance, an English-speaking atheist is not unlikely to exclaim *Oh my God!* when astonished or *Bless you!* when another person sneezes. This does not mean that their espousal of atheism is insincere. Moreover, it is difficult to see how Louw's stance is falsifiable, even in principle. In the BNC, there are eleven instances of *fan the flames of*. Nine of these match the usually noted

negative semantic prosody (with complements such as *popular revolt*, *suspicion*, *misanthropy*, *scandal* and so on). The other two are as follows:

> The work of the modern quantum chemist has helped to **fan the flames** of this debate.
>
> The Tarnished Crown: Crisis in the House of Windsor, is 'perhaps the most significant work ever written on the House of Windsor,' (*Daily Express*, May 8) and promises to '**fan the flames** of royal debate like no other publications since Andrew Morton's *Diana: Her True Story*' (*The Times*, May 3).

Louw's 'binarity' position would lead us to predict that both these cases of debate's flames being fanned *must* either (a) indicate that *debate*, too, is negative; or (b) imply insincerity or irony on the part of the writer. The problem is that both these predictions would rely on unfalsifiable, intuitive personal reactions to the text: either to these concordance lines, or their wider context. This is problematic and is an aspect of a wider problem with reliance on analysis of collocation as a primary explicator of meaning. Louw (2008, 2010) has claimed that collocation is 'instrumentation for meaning' – that is, the study of semantic prosodies via collocation is the necessary 'instrument' for the investigation of what meanings linguistic units have. However, we would argue that this is not strictly accurate. Considered as distributional phenomena, collocation and semantic prosody are not, in fact, 'instrumentation'. The interpretation of collocations as negative or positive in the diagnosis of a semantic prosody is reliant on the linguistic intuition of the analyst (that is, their ability to interpret the meaning of their own language). Likewise, the diagnosis of a usage that runs against a normal semantic prosody as either irony or insincerity is dependent on the analyst's subjective reaction to the co-text, and individual knowledge of the wider context. For instance, Louw's analysis of Clinton *fanning the flames of the peace process* is absolutely dependent on his understanding of the Northern Irish peace process as 'almost as aggressive as the war it is designed to end'. But it cannot be taken as read, simply based on Louw's (non-corpus-informed, intuitive) perception of these events, that there is anything particularly aggressive about the Northern Irish peace process as compared to any other peace process we might care to discuss. So what Louw does here is to import, without comment, a range of historical and sociological assumptions, which others may or may not share, but which are in any case not argued for and not situated in any discussion of social or political theory. Louw is entitled to invoke unfalsifiable opinions in a critical analysis of his reaction to a given text, of course. However, he is not entitled to present an analysis of speaker or writer meaning that relies on them and claim that this constitutes 'instrumentation for meaning' in an objective, narrowly scientific sense. Louw (2000: 60) cites philosopher of science Karl Popper as noting 'that it is the first duty of the scientist to ensure that his claims are potentially falsifiable'. At least insofar as the analysis of the example above is concerned, Louw has signally failed in this duty.

To put it in Hunston's terms of the explanatory versus predictive analysis of semantic prosody: as an explanation of Louw's own reaction to a particular exceptional text, his account of the semantic prosody of *fan the flames of* cannot be faulted. But as a prediction of what another English speaker's reaction must *necessarily* be, or as a prediction of what the author's intention must *necessarily* have been, it is ultimately unsupportable. We see, then, that though semantic prosody as a framework for analysis is undeniably powerful, it is also limited in some important respects.

6.5 Lexis and grammar

6.5.1 The Idiom Principle

As the preceding discussion of collocation and semantic prosody has indicated, the meanings and phraseologies of words in context are major concerns of neo-Firthian corpus linguistics. In fact, some neo-Firthian approaches actually go rather further and make the word, its phraseology and its collocational features the keystone of linguistic description. In the 1980s particularly, corpus lexicography was the central concern of Sinclair and his colleagues (see section 4.4); it is perhaps unsurprising that a school of thought with roots in lexicography should emphasise the absolute centrality of the word – as opposed to some other unit of linguistic description, such as the phoneme, morpheme, phrase or clause – to the understanding of language. A common trend in much neo-Firthian argumentation, particularly regarding grammar, is that many features of language which had traditionally been explained in terms of grammar are instead explained in terms of the lexicon. This ultimately has led some scholars, most notably Hoey, to a position where the distinction between lexis and grammar is effectively abolished – with the grammar being a part of, or an emergent feature of, the lexicon. While most neo-Firthians agree with the privileging of lexis over grammar in linguistic description, this insight has been captured in various ways. Sinclair (1991: 110) expressed it in terms of the 'Idiom Principle':

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments . . .

Sinclair's notion of the Idiom Principle stems from the study of collocation and phraseology. Firstly, it is observable that a large proportion of any running text is made up of fixed or semi-fixed expressions that (a) occur repeatedly in a corpus and (b) have a meaning that is associated with the expression as a whole, rather than with the individual words that make it up. These expressions, Sinclair argues, are produced, perceived and mentally stored as units, not as structures – or in Sinclair's terms, they constitute single lexical choices. To put it another way,

most instances of what may *seem* to be syntactically generated phrases are in fact products of the retrieval of idioms from the lexicon, *not* the product of the operation of a mental set of rules 'on-the fly' (as proposed by many formalist models of grammar in particular). Language produced or interpreted in this way, by the sequential stringing-together of extended multi-word lexical items retrieved whole from the mental lexicon, is described by Sinclair as functioning according to the Idiom Principle. By contrast, in the (relatively rarer) situations where a language user needs to express some meaning for which their lexicon does not contain a pre-constructed idiom, the speaker produces language according to the Open-Choice Principle – that is, by selecting words individually from the lexicon and combining them according to the mental system of grammar rules, in the way that *all* language was traditionally believed to work.

Sinclair's emphases on idioms in the lexicon, and on the Idiom Principle as explaining a far larger proportion of language in use than the Open-Choice Principle, mean that he is effectively proposing a model of language where the lexicon has a much larger role, and the grammar a much smaller one, than in traditional theories of language. However, this view does still invoke the lexicon and the grammar as two contrasting (implicitly separate) systems – necessarily, since a grammar, presumably similar to those proposed by formalist grammarians, is required to account for language structured according to the Open-Choice Principle. Other neo-Firthian thought goes further – not only emphasising the role of lexicon, but also, in fact, *unifying* the description of lexis and grammar largely or entirely. Two systems of linguistic description that are particularly worthy of note for this reason are Pattern Grammar and Lexical Priming.

6.5.2 Pattern Grammar

Pattern Grammar (Francis *et al*. 1996; Hunston *et al*. 1998; Hunston and Francis 1999) is an approach that applies to the study of grammar many of the principles which earlier informed the corpus-based lexicography of the COBUILD school. It does this by seeking to identify the patterns associated with particular lexical items. The *patterns* of a word are defined as 'all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it' (Hunston and Francis 1999: 37). Patterns are identified by means of examining a concordance of a particular word. A phraseology which recurs in such a concordance and which can be described in a schematised way constitutes a pattern. For example, for many English verbs, a concordance will show a repeated phraseology where the verb occurs followed by a noun group (that is, what in traditional grammar is described as a noun phrase). This pattern is described in Hunston and Francis' notation as **V n**. 'v' represents a verb group, 'n' represents a noun group and the 'v' is capitalised to indicate that this is a pattern that centres on the verb. This would cover, for instance, any

Table 6.5 *Hunston and Francis' analysis of interlocking patterns in a sample sentence (central word of each pattern italicised and underlined)*

Sentence: *A fire safety officer said it was important that residents in high-rises were aware of fire safety procedures and equipment in their particular buildings.*

Pattern	Corresponding sentence segment
(no pattern - clause subject)	A fire safety officer
V that	*said* that . . .
it V adj that	it *was* important that . . .
(no pattern - clause subject)	residents
V adj	*were* aware . . .
v-link ADJ of n	were *aware* of fire safety procedures . . .

sentence where a verb is followed by a noun phrase – including both sentences where the noun phrase is traditionally classified as an object, and sentences where the noun phrase is traditionally classified as a subject complement (Hunston and Francis 1999: 151). Another example is the pattern **N to-inf**, which is the form traditionally described as a *to*-infinitive verb complementing a head noun (as in *a desire to do something* or *a decision to do something*: Hunston and Francis 1999: 47).

As these examples illustrate, patterns in general characterise the complementation of the words on which they are centred. However, critically, patterns are not entirely generalised abstractions into which any lexical item could be slotted. Rather, '[p]articular syntactic structures tend to co-occur with particular lexical items, and – the other side of the coin – lexical items seem to occur in a particular range of structures' (Francis 1995 cited in Hunston and Francis 1999: 30), although in some cases words can be classified into 'meaning groups' which share a number of patterns. Furthermore, while a pattern like **V n** can be equated to traditional descriptions of grammatical structure such as *verb + object* or *verb + subject complement*, Hunston and Francis actually argue that such 'structural' analysis is ultimately pointless because '[i]n most cases . . . the structural analysis added nothing, and all that was important to say about a verb could be said in terms of its pattern and its meaning group, irrespective of the structural interpretation' (1999: 152). In Pattern Grammar theory, the production of language proceeds by bringing together lexical items and patterns in interlocking fashion. So, for instance, if a speaker has produced a verb with the **V n** pattern, this can go on to interlock with any pattern that centres on an **N** (for example, **N to-inf** or **N that** or **N on n**), assuming these patterns are compatible with the lexical items being used. An extended example of how patterns flow into one another from Hunston and Francis (1999: 59) is laid out in Table 6.5.

Pattern Grammar can thus potentially explain all language production and comprehension in terms of a single mechanism – that of words and their patterns.

Critically, this single mechanism is centred on *the lexicon* and on meaning. While based on Sinclair's insight of the Idiom Principle, Pattern Grammar therefore goes further and removes the need to posit an entirely separate 'Open-Choice' system of grammatical rules. Hoey's (2005) theory of Lexical Priming shares this distinction, as well as the emphasis on the lexicon, but differs in the details of how language is proposed to work.

6.5.3 Lexical Priming

The starting point for Lexical Priming is Hoey's explanation of collocation. As we have already noted, Hoey considers collocation to be not only a textual phenomenon, but also a psychological phenomenon. To be more precise, he argues that the associations between words that are labelled collocations are primarily a psychological phenomenon, 'the evidence for which can be found statistically in computer corpora' (Hoey 2005: 5). The notion of *priming*, drawn from psychology, is called on to explain the nature of these lexical associations. Priming is the psycholinguistic phenomenon whereby speakers have been observed to understand a given word more quickly if it is presented immediately after a semantically related word, rather than an unrelated word. For example, the word *heart* is understood more quickly (on a microsecond timescale) coming after the word *body* than coming after a word unrelated to human physiology. Hoey argues that just as a person is psychologically 'primed' by hearing the word *body* in a situation such as this, so each word in the lexicon is primed to co-occur with other words (its collocates); thus, when we say a word, we are psychologically primed to subsequently produce its collocates and, likewise, when we perceive a word we are psychologically primed to expect to perceive its collocates afterwards. Thus, the lexical primings create the collocations that are observable in the discourse. At the same time, the lexical primings are also created by our experience of collocation in the discourse. Hoey argues that '[a]s a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context' (Hoey 2005: 8).

Priming is not simply between words; abstractions may also be primed in the same way. So, for instance, it has often been observed that multi-word expressions may have collocations that are distinct from those of the words that make them up; Sinclair's (2004a: 30–5) famous analysis of the expression *naked eye* exemplifies this. Hoey argues that in this case, the words within the multi-word expression are primed to co-occur, and the co-occurring combination is itself primed for particular collocational patterns. Hoey uses the term *nesting* to label this phenomenon. Another abstract form of priming is the possibility for words or combinations of words to be primed with grammatical functions or sentence positions (colligation), or with textual positions or functions (textual colligation). In particular, Hoey argues that a very wide range of grammatical structures and categories may be explained as functions of collocation. For example, even something as basic

as parts-of-speech are argued to be generalisations that emerge from the colligational primings of groups of lexical items (Hoey 2005: 154, 159). The 'noun' category is, in this view, not a primitive linguistic entity. Rather, it is a convenient label for those words which share some particularly common colligational patterns, such as co-occurrence with *the* and *a*; being preceded by adjectives; being followed by prepositions; and so on. Due to the central role assigned to colligation, the account of grammar offered by Lexical Priming encompasses grammar of the type that would be covered by Sinclair's Open-Choice Principle, as well as the grammar of idiomatic language. Thus, like Pattern Grammar, the theory of Lexical Priming removes the centrality of grammar in the description and theory of language, instead assigning the central role to the lexicon and treating grammar as dependent on or emergent from the interaction of items within the lexicon (this point is made explicitly by Hoey 2005: 162).

Notably, Hoey suggests that words (or nested combinations of words) may also be linked via priming to aspects of non-linguistic cognition such as thoughts and feelings – a priming which is two-way; that is, experiencing a given thought or feeling will activate the words it is primed for, just as producing or perceiving a given word will activate the thoughts or feelings it is primed for (Hoey 2005: 161). In this way, the phenomenon of priming is extended to explain not just textual co-occurrence but also the link between signifier and signified, as discussed by linguists from Saussure onwards. Lexical Priming is therefore a comprehensive model of linguistic knowledge, including its relationship to non-linguistic thought. Given this psychological dimension, one of the challenges raised by Lexical Priming theory is how this view of language is to be reconciled with the findings of psycholinguistics. This topic will be addressed further in Chapter 8, but we will give here a single example of the kinds of difficulties that may be observed. Lexical Priming proposes a form of language acquisition that is strongly oriented in conception towards environmental learning, in that it sees mental associations (i.e. primings) being built up through repeated exposure to examples of words in context. However, work by psychologists on language acquisition has shown that the acquisition of words involves factors rather more complicated than simple learning by repeated exposure, although such exposure is indeed one key factor (see, for instance, Tomasello 2003).

In summary, then, a common strand within modern neo-Firthian thought is to build on Sinclair's insight of the Idiom Principle by explaining open-choice language as well as idiomatic language using models of grammar where a central role is assigned to the lexicon and where, in fact, grammar as a mental system is seen as either secondary to or emergent from the lexicon. Both Pattern Grammar and Lexical Priming, in different ways, are attractive for their unifying accounts of lexis and grammar, and for combining the descriptive strengths of traditional neo-Firthian analysis with an interest in the wider question of the underlying nature of language. Of course, the rejection of a separation between lexis and grammar is not unique to neo-Firthian linguistics. It is a common position in the theoretical schools of functional and cognitive linguistics. However, there

is an irony here; although, as we will see in Chapter 8, some neo-Firthians are strictly opposed to the integration of functional-cognitive linguistics and corpus linguistics, their school's success in illuminating the inseparability of lexis and grammar may prove to be a vital factor contributing to this integration. We will return to this topic in Chapters 7 and 8.

6.6 Corpus-as-theory versus corpus-as-method

6.6.1 Data and theory

So far in this chapter, we have surveyed the neo-Firthian approach to corpus linguistics, paying particular attention to this tradition's views on collocation and a particular approach to discourse as definitive of their 'version' of corpus linguistics, to use Teubert's (2005) apt term. In the course of this survey, we have had occasion to point out ways in which this neo-Firthian view differs from other views of corpus linguistics, for example, in terms of the procedure normally adopted to identify collocations. We now have a broad enough understanding of the neo-Firthians' approach to examine one of the central theses of their view of language. This is an idea which, so far as we know, is not held outside the neo-Firthian tradition – it is certainly not a view held by either author of this book. Our treatment of this idea is thus very much an outsider's view.

The idea in question is the proposition that corpus linguistics *has a theoretical status* (Tognini-Bonelli 2001: 1) and may be viewed as being 'an important concept in linguistic theory' (Stubbs 1993: 24). Teubert goes somewhat further than this, identifying corpus linguistics as 'a theoretical approach to the study of language' (Teubert 2005: 2). It has also been claimed that the exclusive use of corpus data itself inevitably leads to the production of new theories of language, in that sense making this approach to using corpus data more 'theory prone' than using it in combination with other methods or with non-corpus-based theories (Römer 2005: 8). The idea that data and theory become in some way co-terminal is sometimes expressed in slightly different terms – such as: the corpus *is* the sole source of a corpus linguist's theory of language – as part of the so-called *corpus-driven* approach (a label we will discuss shortly). Such an approach eschews prior linguistic categorisation, with the implication being that such theories are simply a distraction from what we might term the 'pure' theory hidden within the data (Sinclair 2004a: 191). The strongest forms of the claim imply that the *corpus itself* (and not just corpus linguistics as a field) is the theory; for example Tognini-Bonelli's (2001: 84) assertion that '[t]he theory has no independent existence from the evidence'. This is a very strong stance to adopt and requires some careful unpacking.

We might take this claim that the corpus has a theoretical status as implying a stance on corpus data that is the polar opposite of Chomsky's. The Chomskyan position on corpus data is that it cannot be used as a source of knowledge about the

nature of language. An extreme interpretation of the neo-Firthian position would seem to be, by contrast, that *nothing but* corpus data can be used as a source of knowledge about the nature of language. Or, to put it somewhat facetiously in terms of the physical-sciences metaphor we introduced earlier (section 2.1), if Chomskyans wish to ban astronomy, then (some) neo-Firthians wish to outlaw the dropping of weights from the tops of tall buildings.

However, the neo-Firthian stance would seem to go even further than Chomsky, whose attitude to corpus data is itself notably extreme. In generative grammar, as in most disciplines of linguistics and indeed all the sciences and humanities, there is a clear conceptual division between the theory to be supported (which, in the case of generative grammar, we may assume to be some set of statements about the nature of Universal Grammar) and the evidence intended to support it (a native speaker's judgements of the grammaticality of a certain set of constructed example sentences). If we take what we might characterise as the 'corpus-is-theory' claim literally, then we would be forced to infer that neo-Firthian linguists have erased this data–theory distinction: the corpus, to them, represents at one and the same time the phenomenon in need of explanation and the set of postulates intended to explain it.

Of course, this would be conceptual nonsense. Data is data and theory is theory: observations, while important as necessary support for explanations, clearly do not themselves constitute explanations. We must inquire, then, what is actually *meant* by the claim that the corpus somehow, for some neo-Firthian linguists, takes on theoretical status. It is not hard to understand the roots of this stance. It would seem to be a direct consequence of the centrality of discourse and running text to neo-Firthian ideas – and this, incidentally, may also explain the preference for examining collocation via concordances (samples of texts) rather than via significance calculations which we noted earlier. But while the motivation for the stance may be clear, at the same time it is a priori impossible for text on paper, or even on disk, to explain *itself*. Theorising is a human act which occurs when a human being interacts with data and posits some explanatory principle or principles. So, we may suspect, the claim to believe in 'corpus-as-theory' is actually a claim to be assigning maximal importance to the data in the corpus, a declaration that the analyst from whose interaction with the corpus data the theory emerges will not utilise any theoretical concepts that pre-exist their encounter with the corpus. This interpretation is supported by certain formulations of the 'corpus-as-theory' stance, for instance 'a corpus is not just a tool, but a major concept in linguistic theory' (Stubbs 1997: 301).

Understood in this way, the claim that the corpus, or corpus linguistics, is a theory can be seen as a placeholder for an interconnected complex of precepts and methods that distinguish the neo-Firthian approach to corpus linguistics from the principal other approach, which we will for now refer to as the 'corpus-as-method' tradition. While corpus-as-theory rejects any explanation of language patterns that does not derive directly from the analyst's interaction with the data, corpus-as-method considers corpora and corpus techniques to be sources of empirical data

that may be deployed in support or refutation of any explanatory theory about language – even a theory devised in whole or in part without reference to corpus data.

This is rather hard to understand in the abstract, so let us consider two concrete examples. Firstly, let us return to the enterprise of discourse analysis. As we have noted previously, in recent years a number of studies have begun to apply corpus data and techniques within a CDA-oriented framework. Baker *et al*. (2008: 285) discuss this enterprise as follows:

> theories of language use underpinning CDA result in a focus on grammatical features (e.g., agentivity, passivization, metaphors). The synergy with the particular approach to [corpus linguistics] adopted here adds a focus on lexical patterns. Also, [corpus linguistic] processes can help quantify discoursal phenomena already recognized in CDA; that is, establish their absolute and relative frequencies in the corpus, through the examination of the different linguistic means utilized to express them. Even when the [corpus linguistic] analysis does not set out to examine existing CDA notions, it can utilize a CDA theoretical framework in the interpretation of the findings.

By contrast, Teubert (2005: 4) outlines an approach to 'discourse' in corpus linguistics as follows:

> It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question.

This distinction between these two approaches to the analysis of discourse could hardly be more stark. For practitioners of corpus-as-method, corpus linguistics can be used in interaction with an established analytic framework which may, in and of itself, have nothing to do with corpus linguistics (in this example, CDA). For Teubert, the only appropriate analytic framework for corpus evidence regarding discourse is the corpus-as-theory framework.

To give another example: Deignan (2005), whose work will be discussed in more detail in the following chapter, employs corpus data to address certain problems in Conceptual Metaphor Theory, a branch of cognitive linguistics which explains large classes of metaphorical utterances by reference to a smaller number of mental processes. Again, this is a theory which originated wholly independently from corpus linguistics. As part of her analysis, Deignan identifies patterns of usage which fit within the framework of the theory – but also patterns whose consistency cannot be fully explained by that theory. In this case, we can see that the explanatory theory precedes the corpus analysis, but both informs it and is enriched and refined by it. We will return later to a more detailed consideration of the neo-Firthian perspective on pre-existing theories in corpus linguistics.

This distinction between the neo-Firthian approach of corpus-as-theory and the alternative approach of corpus-as-method has of course been widely noticed in previous work on the underlying ideas of corpus linguistics.[7] One influential and

widely cited characterisation of the split in the field is Elena Tognini-Bonelli's characterisation of *corpus-based* versus *corpus-driven* linguistics, which we have already had occasion to mention. Tognini-Bonelli (2001: 65–6) introduces this distinction as follows:

> the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora become available to inform language study . . . corpus-based linguists adopt a 'confident' stand with respect to the relationship between theory and data in that they bring with them models of language and descriptions which they believe to be fundamentally adequate, they perceive and analyse the corpus through these categories and sieve the data accordingly . . .

On the other hand,

> [i]n a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well-defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus . . . The theory has no independent existence from the evidence and the general methodological path is clear: observation leads to hypothesis leads to generalisation leads to unification in theoretical statement. (Tognini-Bonelli 2001: 84–5)

Contra Tognini-Bonelli, this distinction is slightly more fluid in practice, as such simple binary distinctions often are. A corpus-based researcher may apply a scheme based upon a pre-existing theory but then, when the scheme is applied to data and is found to be deficient, goes on to refine the scheme in what could be termed a corpus-driven fashion. Such a process may be cyclical, as has been well understood by linguists in general and computational linguists in particular for some time. We will return to this point below.

It is clear from Tognini-Bonelli's account of the terms that 'corpus-based' linguistics is to be understood to refer to what we describe here as corpus-linguistics-as-method, and that 'corpus-driven' linguistics likewise should be understood to refer to what we identify as the neo-Firthian, corpus-linguistics-as-theory position. Yet it is not clear that the terms have always been used with this significance. The term *corpus-driven* in particular has been used to refer to any inductive, bottom-up research using raw corpus data, whether or not the research has a commitment to neo-Firthian theoretical positions. Such a definition of *corpus-driven* is offered by Baker (2006: 16), for instance. Biber (2009: 281) claims the label 'corpus-driven' for his approach using lexical bundles, solely on the basis of methodological considerations; but given the nature of much of his other work (see section 5.4), it is clear Biber does *not* subscribe to neo-Firthian theoretical ideas. Meanwhile, Biber's method-oriented definition of *corpus-driven* leads him to argue (2009: 278–80) that some clearly neo-Firthian work (notably Pattern

Grammar) is not in fact wholly corpus-driven. Gries (2010b: 328–9) also treats *corpus-driven* as a methodological term synonymous with 'bottom-up', and, going even further than Biber, argues that all or nearly all self-identified corpus-driven studies cannot be considered 'corpus-driven' in this sense. Clearly while this method-oriented definition has some commonality with Tognini-Bonelli's original definition, it is not by any means synonymous. Yet another perspective on 'corpus-driven' studies is offered by Gilquin and Gries (2009: 10) who say this is characteristic of linguists who 'approach corpus data in an exploratory fashion, i.e. without rigorously formulated hypotheses' which can be (statistically) tested. Furthermore, some authors who object to the corpus-based/corpus-driven distinction have, in response, used 'corpus-based' in a broad sense to encompass *both* approaches as defined by Tognini-Bonelli (e.g. McEnery *et al*. 2006: 11). We have used the term in this broader sense in earlier chapters of this book.

This lack of clarity in the application of the terms *corpus-based* and *corpus-driven* is one reason why, in this chapter, we have elected to use instead the labels corpus-linguistics-as-method and corpus-linguistics-as-theory. However, the main reason is that we believe the *corpus-based/corpus-driven* labels to be ultimately misleading. The implication of corpus-based versus corpus-driven is that the *primary* difference between the two is the degree to which empirical data from a corpus is relied on (and Tognini-Bonelli makes this explicit in her definitions of the terms, as we noted above). But in fact, respect for the empirical evidence of the corpus is probably one of the closest points of agreement between the two traditions of corpus linguistics (consider, for instance, the notion of 'total accountability' to the corpus data, i.e. no data can be ignored as irrelevant; see Leech 1992: 112 and section 1.6.1). Moreover, the corpus-based versus corpus-driven distinction implies a dichotomy where there is actually a sliding scale (an argument also made by Deignan 2005: 88–9). Within what would be dubbed corpus-based linguistics, we see an entire range of roles for the corpus, from providing (at most) a series of examples to illustrate a grammatical theory developed independently of corpus linguistics (Granath and Wherrity's (2005) study of *that*-clauses in English exemplifies this approach) to being the source of most of the claims made (Biber's MD approach is a good example of this). Similarly, studies by practitioners of so-called corpus-driven linguistics do not always rely solely on a corpus in the strict sense: some focus on the language of single texts or single authors ('corpus-driven' stylistics is often of this kind, as, for instance, Stubbs 2005; Mahlberg 2007a, 2007b; so is much of Sinclair's work on text structure, discussed above). So we would argue that the different schools of corpus linguistics are not reliably distinguished – or, alternatively, that their nature is not optimally communicated – by the corpus-based versus corpus-driven distinction, or by the 'top-down' versus 'bottom-up' distinction with which it is often equated. Rather, it is the contrasting stances on the conceptual status of the corpus and of corpus linguistics – as having theoretical status versus as a linguistic methodology – that truly separates the two schools,

although this is only one factor within the division as portrayed by Tognini-Bonelli.

The corpus-as-theory stance does have practical implications for how researchers in this tradition utilise corpus data, however. Sinclair's views on this matter are encapsulated in the phrase 'trust the text'. Used by Sinclair as the title of a volume of his collected papers, this is now effectively a motto for the neo-Firthian school, summing up as it does several of their core ideas: the centrality of discourse, the almost complete exclusion of non-textual evidence, as well as the corpus-as-theory ideology. As a result of these ideas, Sinclair and those who follow him oppose on grounds of principle certain techniques and practices that are commonly used by other kinds of corpus linguist. These include text sampling, corpus annotation and (as mentioned already) the application of corpus evidence within existing theoretical frameworks. Sinclair's concerns with these practices – which, in effect, represent the main thrust of his criticisms of the corpus-linguistics-as-method approach – are stated many times in his writing, but perhaps most succinctly in Sinclair (2004a: 185–93). In the sections that follow, we will argue that Sinclair's position on these issues is not supportable. In some cases, a valid concern is overexaggerated; in others, the very grounds of criticism are ill-founded.

6.6.2 Text sampling

One of the implications of the neo-Firthian exhortation to 'trust the text' is that the integrity of each text within a corpus must be respected. Sinclair (1991: 19) puts it as follows:

> A corpus built up of whole documents is open to a wider range of linguistic studies than a collection of short samples . . .

Later, Sinclair (1996a: 9) was clearer in expressing his disapproval of sample corpora relative to whole-text corpora when he expressed a wish to see whole-text corpora become the default:

> there are still many corpora in use made up of small samples . . . It should, however, be realised that this feature is just a remnant of the early restraints on corpus building and it confers no benefit on the corpus. The use of samples of a constant size gains only a spurious air of scientific method . . .

It is hardly surprising, then, that corpora constructed by Sinclair's group at Birmingham work to the ideal of including only complete texts. This is in contrast to corpora such as the Brown and LOB corpora (and others following the same sampling frame) and the BNC, which are made up of samples of texts (2,000-word samples in corpora like Brown, larger samples – and many shorter texts in their entirety – in the BNC). Part of the reason for the use of samples in these corpora, rather than entire texts, is to maintain, as far as possible, balance

and representativeness (see section 1.4.4). But another, perhaps equally important reason for the use of samples is that copyright permission to distribute a sample of a book within a corpus is generally easier to obtain than the equivalent permission for an entire book, for instance. This is probably one reason why the University of Birmingham's famous, ever-growing monitor corpus, the Bank of English, is less widely available to researchers than are corpora like Brown, LOB and the BNC. It is easy to understand the principled basis of Sinclair's objection, rooted in his long-standing concern for the structure and cohesion of discourse at the level of the complete individual text. Yet it is not so easy to find actual examples of sampling in corpus construction impeding the study of discourse in a corpus made up of samples. We are not aware of any specific analyses, for instance, that a researcher is inhibited from undertaking using the BNC as a result of it containing 'beginning', 'middle' and 'end' samples as opposed to entire texts. The fact is that although a corpus of whole texts offers the *theoretical* prospect of examining the entire discourse structure of each individual text, in *practice* the techniques of analysis on which corpus linguistics is founded typically either are extremely local in scope – such as concordances and collocation – or else abstract away from the textual data, ignoring the *sequential* features of the discourse and merging together the results for different texts – such as frequency lists and keywords. In the former case, it is hard to argue for the necessity of the whole-text sample when individual data points are rarely examined in context greater than a few tens of words at the most. In the latter case, it is difficult to see why a set of whole-text samples has any advantage over a balanced set of beginning, middle and end samples. Given that early-text discourse is different from late-text discourse in detectable ways (clearly shown by, among others, Hoey and O'Donnell 2008), these differences would even themselves out in the merged overall frequency counts for either kind of corpus.

6.6.3 Corpus annotation

A theme returned to from time to time throughout this book has been that of corpus annotation. We must consider it once more here because, over time, the decision as to whether or not to introduce linguistic annotations into a text has become a key fault line between linguists working in the neo-Firthian tradition (corpus-as-theory) and other users of corpus data (corpus-as-method), causing Aarts (2002: 10) to go so far as to state that 'annotation . . . is anathema to corpus-driven linguistics'. In this section we will refer mainly to part-of-speech (POS) tagging as the paradigmatic example of analytic annotation of corpus data (see section 2.3.1). However, Sinclair's criticisms, and the counterarguments we will present here, are applicable in essentially the same terms to other forms of annotation. Sinclair's opposition to corpus annotation appears many times in his writing, in forthright terms usually similar to these:

> Of course, one cosy consequence of using tagged text is that the description which produces the tags in the first place is not challenged – it is protected. The corpus data can only be observed through the tags; that is to say, anything the tags are not sensitive to will be missed.
>
> And, ultimately, as a side-effect, text becomes grossly overstuffed with tags, and processing speed is affected. [. . .] In corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then the patterns of this uncontaminated text are able to be observed . . .
>
> (Sinclair 2004a: 191)

Sinclair (2004b) modifies that position slightly, accepting that some annotations, namely those that are for specific research purposes, may be acceptable. But he sees those linguistic annotations that are made available as part of general-purpose corpora, such as POS tags in corpora like the BNC, as particularly problematic: '[t]heir inclusion among generic resources . . . is misplaced and hazardous, and it holds back progress substantially' (Sinclair 2004b: 54).

We have already discussed some of these objections to corpus annotation earlier in this book (section 1.5); let us here summarise, and give some critical commentary. Sinclair's main objections to the practice of tagging can be placed under two main headings: (a) tagging represents a violation of the integrity of the text; (b) tagging imports non-corpus-motivated frameworks of analysis, interpretation of the corpus data thus being limited by these frameworks. The latter objection – which is in substance identical to the criticisms made by Tognini-Bonelli of 'corpus-based' linguistics – is more serious, and will be dealt with in the following section. However, objection (a) is fundamentally ill-founded.

Sinclair argues that the insertion of tags into running text causes the text to lose its integrity. This concern too, then, stems ultimately from the imperative to 'trust the text'; and, superficially, it is no doubt true. The linguist's experience of

> The_ART cat_N sat_V on_PREP the_ART mat_N

is no doubt different from the experience of reading

> The cat sat on the mat

But to condemn tagging for this reason is farcical. Most modern concordancing software makes it possible to hide tags when viewing a concordance, especially when XML encoding is used:

```
<w pos="ART">The</w> <w pos="N">cat</w>
<w pos="V">sat</w> <w pos="PREP">on</w>
<w pos="ART">the</w> <w pos= "N">mat</w>
```

as is standard and recommended practice (Burnard 2005) in twenty-first century corpus construction. Even the adjustments to the tokenisation made by some POS taggers (e.g. splitting negative forms like *don't* into *do n't*, or merging *of course* into *of_course*) are undone in the rendering of concordances by particularly

powerful software such as BNCweb (Hoffman *et al*. 2008). More fundamentally, the process of tagging a text does not delete the source plain text. It can be, and frequently has been, retained and distributed alongside the tagged version, contrary to Sinclair's apparent claim (2004a: 191):

> The interspersing of tags in a language texts is a perilous activity, because the text thereby loses its integrity, and no matter how careful one is the original text cannot be retrieved. Thankfully it is no longer necessary to mix the two except for specific moments of application, and in The Bank of English, for example, the tag strings are always kept apart from the text itself in parallel data streams.

Sinclair's fundamental error here is the confusion of form with substance. Regardless of whether annotation is embedded in the text or stored in stand-off files, it may be rendered or not rendered within concordance lines according to the user's preference, given a reasonably state-of-the-art concordancer. To claim that a text with tags interspersed within it cannot be recovered reliably is clearly untrue, at least where a systematic encoding scheme is used. The difference between embedded and stand-off annotation is purely a matter of implementation. It has no substantial impact on the analysis of a text. A criticism of embedded tagging because it is not stored in the stand-off format is clearly a criticism that lacks any substance.

Considered in a wider context, the claim that tags disrupt the integrity of the text may appear to be so extreme as to be risible. This book was stored as text in a word-processor file that contains text-formatting markup. Programs interpreted that text and displayed it on a screen, and printed it to this page, in a way that masked the markup. Did that process impair the integrity of this text for the reader? Of course not. It is consequently difficult to argue that the presence of tags in the underlying disk storage of a text – or even in the rendered concordance line – is any more of a violation of the integrity of the text. It is certainly less of a violation than the simple act of creating a concordance. Running a concordance search means pulling the words out of the context of their original publication, displaying them with a mere dozen or two dozen words of the co-text they originally possessed, and laying them out vertically alongside other ripped-out snippets which merely happen to contain another instance of the same word. All this is surely as great a violation of what Carter (in Sinclair 2004a: 5) calls Sinclair's 'major principle of respecting and trusting the integrity of the complete text as the basis for linguistic description, analysis and theory building' as is the application of grammatical labelling. More importantly, the common process of deleting pictures, tables and other non-paragraph material from corpus texts is clearly a much grosser violation of the text than introducing annotation can conceivably be. Yet this is a violation that a great majority of corpora, including the Bank of English, have carried out and will in all likelihood continue to carry out. Given that such deleted items may materially alter the interpretation of the remaining text, it is curious that such a violation has been overlooked.

Setting aside the issue of stripping non-textual material from corpus texts, we do not seriously wish to argue that there is anything wrong with concordancing. Rather, our point is that the fundamental nature of all corpus methods is that they sample, summarise, quantify, group together and abstract away from the sentences of the corpus texts as experienced by their original readers. So it does not seem supportable to object to the annotation of corpora on this ground alone.

Sinclair's final criticism under this heading is that the insertion of tags into a text may have the adverse effect of slowing down corpus searches. On this point, little need be said. It may have been a valid concern in the early 1990s (although even then the issue could be addressed, as Sinclair himself notes, by the use of stand-off annotation files, which differ from in-text tags only in form and not in substance). However, for many years the power of the average desktop computer has increased much more quickly than the size of the corpora that linguists wish to analyse. As long as this trend continues, Sinclair's concern is simply not a valid criticism.

However, there are many excellent reasons to be wary when exploiting POS tagging. Not least among them is the fact that even the best tagging software struggles to achieve much more than 97 per cent accuracy on unrestricted text. It is also clear that decisions made according to any scheme of linguistic description necessarily entail some relatively subjective decisions about the assignment of instances into classes (see section 2.3.2). Analyses of this sort, whether they are explicit or not, are always bound to be linked to a degree of controversy. But the groundless concerns expressed by Sinclair with regard to tags disrupting the integrity of the text, or causing corpus processing to slow down, are not among these valid reasons for caution in the use of tags.

At the core of the objection to annotation is, in fact, the objection to the application of pre-existing theories and categories to corpus data which we have already mentioned as an important aspect of neo-Firthian thought. This is where the true fault line lies; it is simply *reflected* in the discussion of corpus annotation, as in this comment from Sinclair on the *Longman Grammar of Spoken and Written English* (Biber *et al*. 1999):

> The painstaking efforts and academic honesty of Biber *et al*. (1999) is worth noting here, because they doggedly follow the model of Quirk *et al*. (1985) and so they do not have a chance of aligning their received categories with the evidence from their corpus. So they resort to talk of the 'nouny'-ness of nouns . . . and are most unconvincing in their attempt . . . to hold onto 'species nouns' like *sort*, *loads* as still in the same class as nouns like *boy* and *bicycle* . . . (Sinclair 2004b: 53)

Sinclair (2004b: 56) concludes his analysis of the *Longman Grammar* thus:

> This grammar explicitly applies a pre-corpus model of language to a small corpus and annotates the corpus as a first step. Despite what must have been an enormous effort of silent editing, the evidence that surfaces in the book consistently fails to validate the categories of the imposed description . . .

We can thus see that Sinclair's view on annotation is, fundamentally, an incidental complaint motivated by a deeper objection to 'pre-corpus' models. Accordingly, in the next section we will continue to explore this issue at the heart of the matter.

6.6.4 Corpus linguistics and pre-existing theory

The second set of objections to corpus annotation made by neo-Firthians are rather more serious and deserving of more detailed consideration. This is that any given annotation scheme necessarily represents a theoretical framework that is non-corpus-based, since the categories within the scheme must be devised *prior* to the scheme's application to text. The application of such annotations, then, restricts all analysis based on the annotated corpus to operating within the parameters of the theoretical framework underlying the annotation scheme.

This is allied to the neo-Firthian view of using corpora and corpus methods to provide empirical evidence within an existing theory or framework. As we have seen in the discussion of corpus-as-theory and the corpus-based/corpus-driven distinction, any such approach is rejected by the neo-Firthians. The corpus *is* the theory; the linguist's interaction with the corpus is the only legitimate source of explanatory generalisations. Both in the case of annotation and in the case of applying corpora within other disciplines of linguistics, the neo-Firthian accusation is that the corpus-as-method approach undermines the power of the corpus to instantiate its own analytic categories and to inspire its own theoretical constructs, with Sinclair (2004b: 53) claiming that, for scholars working with a particular theory in the corpus-as-method approach:

> [t]here is no impetus to expose the theory to scrutiny. Overwhelmingly the consensus view of researchers is that the models are basically correct, and while they can be tidied up by corpus evidence there is no need to open up the whole complexity of language theory and description for the sake of some minor blemishes. Better to get on with the job.
>
> In the view of the corpus-driven linguists, the picture is quite different.

Again we see the underlying theme: trust the text, distrust established theory. A similar charge is levelled at the corpus-as-method approach by Tognini-Bonelli (2001: 67–71) in her argument that corpus-as-method approaches imply 'insulation' of the data from the theory, that is, that 'in this approach the data is relegated to a secondary position with respect to the theoretical statement proper' (2001: 68). These are serious accusations which, if valid, would mean that much of the work undertaken within the corpus-as-method tradition is fundamentally futile. However, there are a number of grounds on which, we will argue, these criticisms of corpus-as-method fall down.

Let us first consider the issue of circular analysis – that is, the situation where a non-corpus-motivated theoretical framework forms the basis of an annotation system and thus ensures that the analysis of the corpus that uses the annotation

cannot but confirm the theoretical framework underlying the annotation. This is a powerful criticism in principle. However, there are two reasons why it is not, in practice, a critical concern. The first relates to the nature of the corpus-as-method approach. It is obviously *possible* to derive an analytical scheme from a theory which was not developed using corpus data; this is trivially true and is hardly worth discussing here. What it is not possible to do is to shoehorn corpus data into an analytical scheme that is fundamentally flawed. The exposure to corpus data of an analytical scheme based on a non-corpus-informed theory is a critical test of that theory, not an uncritical reconfirmation of it. We will give two examples from our own research.

It is a very common experience in research on corpus annotation for the annotator to find that an initial annotation scheme – which may be based on intuition or on the non-corpus-linguistic literature – proves inadequate, and must be revised, once the attempt is made to apply it to the data. For example, Hughes (1998) describes eight categories of usage for swearwords, established on the basis of introspection or isolated examples. In a corpus-based study of bad language, McEnery *et al*. (2000a, 2000b) attempted to use Hughes' system as an annotation scheme for instances of swearwords extracted from the BNC. However, it proved impossible to adequately classify all examples of swearing by the use of this scheme. Some categories were observed to include very different usages which needed to be categorised separately. In other cases, instances in the data did not fit *any* of the original categories. The extensions to the scheme resulted in a final total of sixteen different categories; of those categories surviving from the initial scheme taken from Hughes (1998), many were defined differently (see McEnery *et al*. 2000b: 45).

Similar experiences are common in POS tagging as well. Hardie *et al*. (2009) outline the development of a tagset for Nepali, a language which has been the subject of much grammatical analysis but was not formerly POS-tagged. Initially, single tags were proposed for nominative, accusative, genitive and ergative case nouns, as these are the cases most often cited for Nepali. However, attempting to apply these tags in actual text quickly demonstrated a variety of configurations of case morphology which this scheme was incapable of covering adequately. An entirely different approach to case markers in Nepali – treating them as clitics which can be tagged as independent tokens from the bases they are attached to – was required to account for the data in a tractable way. Of course, as with the swearword analysis, it was a commitment to total accountability to the corpus data that necessitated an analytic scheme capable of covering *everything* encountered in the corpus, rather than just a subset.

Quite simply, then, as corpus methodologists working intensively to apply analytic schemes developed from non-corpus-informed theory, it is our impression that it simply is *not possible* to apply an annotation system to a corpus if it is fundamentally incompatible with the patterns of language use that the corpus contains. In our view, then, Sinclair's and Tognini-Bonelli's concerns regarding analyses that do not reject all non-corpus-derived frameworks are simply not

supported by actual experience. Further evidence for this view can be drawn from the very large number of cases in which some corpus analysis, undertaken within a given analytical framework or given theory, has not simply reconfirmed but has actually challenged, extended or even disproven one or more points of that theory. Let us take for example Pullum and Scholz (2002). In this study, the basic framework of Chomskyan grammatical analysis is adopted. In particular, the focus is on the phenomenon of structure dependency, as exemplified in English interrogative clauses containing embedded clauses, such as the following:

> The man [who is insane] is smiling.
> Is the man [who is insane] ___ smiling?

This type of sentence is the object of search in Pullum and Scholz's study. Effectively, then, they adopt the theoretical assumptions of generative syntax – since the analysis of English interrogatives as involving movement from an underlying declarative structure is not found in all theories of grammar. So this is a very clear example of a pre-corpus or non-corpus theoretical framework – indeed generative syntax is, of all linguistic theories, probably the most inimical to corpus research. But the *result* of this research was not a reconfirmation of generative theory. It was quite the opposite. Pullum and Scholz – by finding many example sentences of exactly this type in a range of different text corpora – refute a claim that such sentences are incredibly rare: Chomsky has even said that a person might go their entire life without being exposed to such examples (Piattelli-Palmarini 1980: 40). This is not an incidental claim; the rarity of such sentences is proposed by Chomskyan linguists as evidence for the unlearnability of this syntactic structure, which in turn is used as an argument for the innateness of many key aspects of grammar.

So here we have a clear example of non-corpus-motivated theory informing – indeed, determining – the approach taken to the corpus data. However, far from reconfirming the theory in question, this study in fact *disconfirmed* a thesis of fundamental importance to the theory. Admittedly, in this case, the very purpose of using a Chomskyan analysis to determine what to search for in the corpus was precisely *in order* to argue against Chomsky's view of language acquisition. Even so, this remains in our view a clear example of the sort of thing that should not exist at all if Sinclair's criticism of the corpus-as-method approach were correct: namely, a theory-informed, 'corpus-based' study coming to a conclusion *directly contradictory to the theory that informed the analysis*.

There are many other cases where a corpus study shaped by a particular theory has led to the underlying theory being challenged, modified or extended; some such research will be reviewed in the following chapter. So, the collective experience of methodologically oriented corpus linguists has been that, far from limiting corpus analysis to reflecting and reconfirming the analyst's original assumptions, corpus-methodological studies can valuably add to or amend the theoretical framework on which they were based, whether that framework was originally corpus-derived or not. This is equally true whether the framework enters

the study in the form of the linguist's prior knowledge and expectations, or as the basis of a system of annotation. This has long been accepted by the corpus-as-method approach, as the following quote from Jan Aarts in 1988 (reported in Aarts 1991: 45–6) shows. He is discussing the process involved in the development of a grammar:

> The (corpus) linguist first writes a formal grammar for some well-delimited part of the corpus language, for example, the noun phrase. This grammar is written on the basis of the linguist's intuitive knowledge of the language and whatever is helpful in the literature. The first version of the grammar is then tested on a set of sentences made up by the linguist himself. On the basis of the analysis results the grammar is revised, tested again, and this process is repeated until the linguist has the conviction – or perhaps better, the illusion – that his grammar is tolerably reliable and complete. At that moment the grammar is tested on the sentences of a corpus – which is always a dismaying experience. Only linguists who use corpus data themselves will know that a corpus always yields a greater variety of constructions than one can either find in the literature or think up oneself.

The typical neo-Firthian response to such defences against Sinclair's criticisms is to discount them as insufficient. For instance, Tognini-Bonelli argues that in cases where, in 'corpus-based' linguistics, analytic categories are 'modified by confrontation with corpus evidence' there is 'an implication that any modification will be of a minor nature' (2001: 74); in consequence, 'the commitment to the data as a whole is ultimately not very strict or systematic' because 'the relationship between an item and its context is not taken as systematic and determining in the definition of linguistic categories' (2001: 81, 86). But we would argue that, while we may or may not subscribe to the exact process that Aarts uses to develop a grammar, the spirit in which he engages with the data, realising that it will challenge and revise the analytic scheme with which he approaches it, is much closer to the spirit in which corpus-as-method linguists operate than Sinclair, Tognini-Bonelli and some other neo-Firthians' caricature of such work as an uncritical importation of sacrosanct, insulated theories that will be preserved regardless of the data. Corpus-as-method researchers do *not* approach the corpus in the cosy anticipation that they will not find their analytic framework or supporting theory challenged. They approach it in the full expectation of a 'dismaying experience'. Accordingly, we would argue that the neo-Firthian criticism of the corpus-as-method approach is, in this respect, unfounded.

Finally, we may ask whether the approach of corpus-as-theory, and Sinclair's associated dictum to 'trust the text', are in fact appropriate as a conceptual framework for neo-Firthian corpus linguistics. As noted earlier, 'corpus is theory' should almost certainly be read as a shorthand for 'the interaction of the linguist with evidence drawn directly from the corpus is the only legitimate source of linguistic theory'. And in this context, one of the implications of 'trust the text' is that 'the linguist should approach the text with no preconceptions'. This is implicit

in the neo-Firthian rejection of prior theory as a basis for corpus analysis, as discussed above. However, it is arguably impossible to approach corpus evidence with no preconceptions about language. Let us once more take part-of-speech classes as an example. Sinclair's analyses, along with those of many of his colleagues, very frequently refer to such conventional categories as noun, verb, adjective and so on. For example, Sinclair's (1991: 82–3) discussion of the word *of* assumes the existence (and defining features) of a 'preposition' category in the course of arguing that *of* is not a member of this category; related concepts such as a 'noun' category and the idea of a 'head' noun are also assumed in the course of the subsequent analysis (Sinclair 1991: 85–94). More extended analyses of grammar, such as Pattern Grammar (see section 6.5.2), likewise assume these categories, and the 'groups' of words found around them, as part of the descriptive apparatus.

It is unsurprising that this should be so. The ability to identify such conventional categories in a sentence is a part of the analytical competence that a linguist brings to the study of the concordance. Yet these are the very kinds of categories which, in a strict neo-Firthian approach, should be derived from the data rather than imported from prior metalinguistic knowledge. Of course, categories such as noun and verb *can* be – and have been – derived directly from corpus evidence (see, for instance, Schütze 1995). But this does not obviate the plain fact that, by the time any budding linguist approaches their first concordance, they will already have been inducted – at least once and more probably many times – into the hallowed and venerable theoretical framework of nouns, verbs, adjectives, adverbs and all the rest. Put bluntly, no one approaches corpus data with no preconceptions whatever. So the involvement of pre-existing theoretical frameworks in corpus analysis becomes a matter of degree. All analysts bring some prior understanding of language to the analysis; this is unavoidable and, in all likelihood, useful. Intuition can be an extremely good guide in approaching a corpus. Put simply, a linguist's intuitions can provide a framework of informed guesswork allowing them to interact with a corpus rapidly and effectively. Likewise, pre-existing theoretical frameworks can assist the corpus linguist in a similar way.

Since prior understanding cannot be eliminated, the question is *how much* that prior understanding is relied on rather than *whether* that prior understanding is relied on. The dedicated neo-Firthian may take care to assume as little as possible in advance; the functional grammarian seeking corpus support for a position originally arrived at for entirely different reasons may, by contrast, assume a great deal and use those assumptions deliberately and explicitly to approach the analysis. But this is a matter of degree, not kind. The distinction between the two approaches to corpus data is quantitative, not qualitative.

So, just as above we argued that the corpus-based versus corpus-driven distinction is ultimately unhelpful – because the level of emphasis placed on corpus evidence is a matter of degree – so too is the extent to which theoretical concepts are with the linguist prior to their encounter with the data, as opposed to emerging in interaction with the data. The emphasis in the neo-Firthian tradition on

trusting the text and corpus-as-theory has, we would argue, thus served to obscure the extent to which neo-Firthian corpus linguists and linguists of various other stripes are engaged in the same enterprise. They certainly come up, frequently, with similar results. For instance, it may be because of the emphasis on the principle of the corpus as the sole legitimate source of knowledge about language that many neo-Firthians have not realised the massive convergence of their own findings with various forms of functional and cognitive linguistics, particularly in the matter of the growing consensus, variously phrased, of the inseparability in theory and in practice of lexis and grammar. As noted in section 4.4, Sinclair and his colleagues have contributed immensely to the contemporary understanding of how lexis and grammar are interrelated. But by rejecting all other theoretical stances, some neo-Firthians run the risk of isolating themselves from this convergence.

In the end, then, there can be seen to be some inherent problems in 'trusting the text', as Sinclair's slogan has it – at least according to the understanding of that principle espoused by some contemporary neo-Firthian scholars. As we have argued, to the extent that 'trust the text' is interpreted as something like 'exclude any sort of linguistic explanation other than that which emerges from the interaction of a linguist with corpus data, which must be approached without the use of any prior linguistic theory', it is a dictum whose necessity is in practice unsupportable and which cannot, as a matter of fact, be followed. To the extent that 'trust the text' is interpreted as 'do not disregard empirical evidence from language in use', it is hard to imagine any linguist other than perhaps a Chomskyan formalist disagreeing with it: seen thus, the injunction to trust the text is banal.

6.7 Summary: Sinclair's contribution to corpus linguistics

There is little doubt that, among the neo-Firthians, the work of Sinclair has been deeply influential. But his influence extends beyond that tradition as well. He was the founder and long-term leader of the work of a dedicated team of corpus linguists at the University of Birmingham, which has been, and remains, a major centre in the development of corpus linguistics in the UK and around the world. Furthermore, his contributions to empirical research on collocation, and above all to lexicography, are of immeasurable value.

Yet the limitations of Sinclair's approach to corpus linguistics may be to some extent adverse consequences of its strengths. The extensive study of collocation is a primary strength of his research in particular and neo-Firthian research in general. But the over-extension of claims regarding, in particular, the role of collocation in meaning can lead to some deeply problematic statements being made by neo-Firthians. For example, in a discussion of words taking on meaning in the context of co-occurrence with other words, Sinclair (1996b: 82) comments that 'the idea of a word carrying meaning on its own [can] be relegated to the

margins of linguistic interest, in the enumeration of flora and fauna for example'. This surely goes too far. Consider the fact that, in the BNC, the word *house* collocates with the word *build*. It would not seem to us especially radical to argue that this collocation is an effect of, rather than a cause of, the meanings of the words in question. The entities that *house* refers to are, out there in the actual world beyond the discourse, very frequently associated with the actions that *build* refers to. The collocation of the words thus seems to be a result of their semantics, and not merely a matter of idiomatic convention. By contrast, the fact that *house* collocates with *build* much more strongly than with the near-synonym *construct* (as measured by mutual information) is a fact about English phraseology, and is arbitrary relative to the core semantics of the words. Sinclair's disregard for aspects of lexical semantics other than collocation risks obscuring such potential distinctions between collocations that are linguistic phenomena per se and collocations that can be explained as epiphenomena. So the overwhelming role that Sinclair assigns to collocation, to the almost total exclusion of non-collocation-based meaning, is surely a weakness.

Likewise, Sinclair and his colleagues' focus on discourse has explained much about the structure of language at the textual level. But an over-focus upon the centrality of text and discourse – as embodied in the claim that the corpus can be or should be a theory of language and that no other theory should be entertained – is a problem for the neo-Firthians, as explored in this chapter. It led to the counterproductive distinction between corpus-based and corpus-driven linguistics, the mantra of 'trust the text', and the elevation of methodological preferences (such as the attitude taken to corpus annotation) to the level of theoretical precepts. All of these are problematic to differing degrees. In proposing such an extreme form of neo-Firthianism, Sinclair may have succeeded in defining the boundaries of a distinctive approach to language, but he simultaneously outlined the *limitations* of the neo-Firthian approach and raised needless barriers between neo-Firthian and non-neo-Firthian linguists.

It is these aspects of Sinclair's thought which, we have argued in this chapter, are ultimately unsupportable, or in the worst case, banal. We should emphasise at this point that it is Sinclair's stance on these *theoretical* issues which we wish to criticise. In terms of the discoveries made about English, the methods, the insights, the *actual results* he produced, we find far less to disagree with. It must also be emphasised that not all neo-Firthians concur with all aspects of Sinclair's position. For instance, as we have noted, Hoey's theory of Lexical Priming moves beyond the corpus-as-theory position in several important respects, while still maintaining the focus on collocation and the structure of text and discourse that are central to neo-Firthian corpus linguistics.

The link between corpus and theory will be picked up again in the next chapter, where we will explore an example of how theory and corpora can be gainfully combined by looking at the intersection of functional and cognitive linguistics with corpus linguistics. Not only does this provide a compelling example of how the corpus-as-method approach can stimulate theoretical innovation, it also

shows how, through this process, results may be arrived at which are not entirely incompatible with the corpus-as-theory approach (a topic we will address in Chapter 8). However, arguably such results can be achieved much more swiftly when existing theory is adopted and adapted rather than being discarded and reinvented.

Further reading

Many of the works we have cited in this chapter as statements of the neo-Firthian position are also textbooks that give overviews of corpus linguistics as a whole and, in some cases, examples of how neo-Firthian analyses are to be carried out. If you are interested in the theoretical debates we have discussed in this chapter, we strongly recommend reading one or more of these works, which include Tognini-Bonelli (2001), Stubbs (1995), Stubbs (2001) and Teubert and Čermáková (2004; this work covers more-or-less the same arguments as Teubert 2005). Stubbs (2001: chapter 10) in particular has some thoughtful comments on theoretical issues. Similarly, for a more in-depth view of neo-Firthian grammatical theory, Hoey (2005) on Lexical Priming and Hunston and Francis (1999) on Pattern Grammar are strongly recommended. Finally, the single work which perhaps encapsulates neo-Firthian corpus linguistics best is Sinclair (1991), which is in addition a very accessible and enjoyable read; after that, Sinclair (2004a) is the work of Sinclair's which covers most ground concerning his view of linguistics.

Practical activities

(A6-1) Choose a relatively frequent noun (at least 25 per million words; if you are working with a relatively small corpus, make sure you have at least a couple of hundred instances to look at). Produce a concordance of the noun and examine it for potential collocations without using any statistical tools. You may need to look at several different ways of sorting the concordance (one-word-left, two-words-right, and so on).

- What words does your noun collocate with?
- Do any of those words fall into groups that might suggest one or more semantic preferences?
- Are there any repeating grammatical patterns around your noun that might suggest colligations of any kind?

(A6-2) Now use your concordancer to produce a list of statistical collations for the same noun (preferably based on one of the commonly used statistics such as log-likelihood, mutual information, t-score or z-score). Are all the collocates you identified by manual analysis of concordance lines also pulled out by the automated procedure? Does the automated procedure pull out any collocates that you *didn't* spot? How about the semantic preferences and colligations – are they evident from the statistical list of collocates or not?

(A6-3) The neo-Firthian notion of the (extended) unit of meaning suggests that polysemous words are, in fact, never semantically ambiguous in context, because there will always be some element of the context that disambiguates them, showing clearly which sense is meant. Choose two or three highly polysemous (or homonymous) words, such as *run*, *bank*, *save*, *skip* or *try*.

- Think of as many different senses as you can for each word.
- Look them up in a dictionary – it is best not to use a really big dictionary or else the entry will be too long to read easily! Did you miss any senses? How many senses are there in total, according to your dictionary?
- Now, for each word, do a concordance and thin it to a random selection of fifteen or twenty concordance lines. Look carefully at each example.
- What sense of the word is active in each example?
- Is there a specific collocation evident in the concordance line which indicates the active sense?
- If there is not a specific collocate that disambiguates, consider the grammatical context – whether it is being used as a noun or a verb; whether the verb is transitive or intransitive; and so on – does this disambiguate the sense?
- Did your corpus searches find any genuinely ambiguous examples? If so, how much of the extended context do you have to read to work out which sense is meant?
- Did your corpus searches find any examples of senses *not* covered by the dictionary you consulted? (This is not likely, but *might* happen.)

Questions for discussion

(Q6-1) See if you can find examples of 'pre-corpus' and 'post-corpus' dictionaries – ideally, earlier and later editions of the same dictionary, but any pair will do as long as one is based on corpus analysis and the other isn't. Choose some words that are relatively basic vocabulary (e.g. *receive*, *travel*, *miss*, *visit*), look them up in the dictionaries, and compare the information each dictionary contains and how that information is presented. What differences can you see? How has corpus analysis informed the creation of the dictionary entry? Has this improved the quality of the dictionary entry – if so, in what way?

(Q6-2) (a) Some corpus tools allow collocation to be studied based on lemmata. Others only allow collocation statistics to be extracted based on wordform searches. What do you think might be the pros and cons of working with lemmata rather than wordforms when studying collocation?
(b) In neo-Firthian analyses, in particular, collocation based on wordforms is the standard, and collocation based on lemmata is generally avoided. Why do you think might this be?

(Q6-3) In section 6.2 we discussed an approach to collocation that only considers two words as collocating with one another if they occur in a specified grammatical relationship to one another – for instance, verb and direct object; or head noun and premodifying adjective. In this view of collocation, simple co-occurrence of two words in proximity is *not* enough. What theoretical commitments are inherent in this view of collocation? Are these assumptions compatible with neo-Firthian theory? Why, or why not?

7 Corpus methods and functionalist linguistics

7.1 Introduction

In this chapter, we will show that an important trend in contemporary corpus linguistics is the coming-together of the analytic techniques, theories and findings of corpus linguists on the one hand, and on the other, of theoretical linguists of the school sometimes labelled as *functionalist* (or *cognitive*, or *usage-based*: see discussion in section 7.2 below); in the next chapter, we go on to develop a similar argument with regard to experimental psycholinguistics.

As the previous chapter illustrated, within corpus linguistics there is a distinction between two schools of thought which we have dubbed 'neo-Firthian' or 'corpus-as-theory' and by contrast, 'corpus-as-method' or 'methodologist'. Across this chapter and the next, we will argue that the rapprochement between corpus linguistics and functional linguistics is evident in both the neo-Firthian and the methodologist traditions, but that it takes a rather different form in each. In the methodologist school, it has two manifestations. Firstly, functionalists have gradually been taking more and more regard of the importance of corpus evidence. For linguists working in a 'usage-based' paradigm, as we will see, a massive collection of 'usage' can be invaluable. Conversely, some corpus linguists – including, notably, Stefan Gries[1] – have used functional and cognitive linguistics as a theoretical framework within which to interpret the outcomes of corpus analysis and from which to motivate new techniques for the investigation of corpus data. The rapprochement of the neo-Firthian school of corpus linguistics with functional-cognitive linguistics and psycholinguistics will be the topic of Chapter 8.

This chapter is, therefore, structured as follows. Firstly, we will give a brief overview of the context and theoretical frameworks of the several varieties of theoretical linguistics we are grouping together here under the broad covering title of 'functionalism', including cognitive, functional, typological and usage-based approaches to language. We then go on to describe, in sections 7.3 to 7.5, the interface between functionalism and corpus methodologies. In doing so, we will review not only research in functionalist linguistics that has utilised corpus methods, but also research in corpus linguistics that has incorporated functionalist theoretical frameworks.

7.2 Functionalism in linguistics: a brief overview

What does *functionalism* mean in the context of linguistics? In the broadest sense, functionalism is one of the two main schools of thought within linguistic theory as it has developed since the 1950s. Functionalism is so called in contrast to the formalist school of linguistics, of which the primary exemplar is Chomsky and the generative approach to language. Its distinguishing features are most easily outlined in terms of what it is *not*. Formalist linguistics analyses language form *in isolation* not in context. That is, knowledge of language is conceived of as an abstract system which is to be explained entirely in its own terms. It is not to be explained by any aspect of general, non-linguistic cognition. From this precept stem many other aspects of Chomskyan theory: the distinction between competence and performance, the rejection of corpus data and reliance on introspection, and the view of language as an autonomous cognitive system. Functionalism, in a nutshell, is the rejection of this precept: functionalists investigate language form, but explain it with reference to the functions to which language is put. Language is not seen as an abstract, isolated system, but one that is *used* to communicate meaning, and which is shaped by the ways it is used, by the contexts in which it occurs and by the structure of human cognition. 'Functionalism' in this broad sense covers a set of approaches to the theory of language sharing these features, including functional linguistics, cognitive linguistics and language typology. The emphasis on language *in use* makes functionalism compatible with corpus linguistics in a way that formalist linguistics is not.

This distinction is not absolute. For instance, formalist linguists of the Chomskyan school are willing to call on corpus data in some circumstances – most notably in the study of early child language, where the intuition of an informant can never be available, because children develop metalinguistic awareness rather later than language itself. For example, Bloom (1990) uses child language data, in the form of verb-phrase length in sentences with and without subjects, to argue that English-speaking children omit subjects not because they are unaware that the subject is compulsory, but because the sentences with subjects are too long for them to cope with – that is, due to performance limitations. This suggests, in generative terms, that the 'pro-drop parameter', the switch controlling whether or not subjects are compulsory in a particular language, has the correct setting from the outset. Other studies which have addressed issues in early grammar using real performance data – mostly exploiting the CHILDES database of child language corpora (MacWhinney 2000), which we will discuss in detail in section 8.2.2 – include Déprez and Pierce (1993); Hyams and Wexler (1993); and Snyder and Stromswold (1997). Nevertheless, functionalism is, in general, much more open than formalism to the use of corpus data and corpus methods.

It is, of course, an oversimplification to treat cognitive, functional and typological linguistics as synonymous. They are, rather, distinct but overlapping approaches to the study of the nature of language. Where they coincide is in

their concern with the nature of language and its relationship to the nature of thought; that is, they are strands of what is usually termed *theoretical* linguistics as opposed to *applied* linguistics. They are also alike in their opposition to the Chomskyan view that language can (and should) be studied in isolation from usage and cognition.

Cognitive linguistics approaches language in a way that prioritises human thought as an explanation for the observed characteristics of language. In particular *conceptualisation*, or how people construct abstract concepts and schemata to think about the world, is a major explanatory factor. More specifically, language is explained in terms of modes of cognition that are domain-general – not particular to language – and that can be demonstrated to exist independently of their role in language, for example through psychological experimentation. There is a variety of strands of research within cognitive linguistics (Langacker 2008: 7). Most are concerned with either semantics or syntax or, perhaps most characteristically, the interaction between the two. There are comprehensive theories of grammar rooted in cognitive linguistics, most notably Cognitive Grammar (Langacker 1987, 1991, 2008) and Construction Grammar (Goldberg 1995; Croft 2001); other topics of wide interest include the impact on language (and semantics in particular) of cognitive processes such as categorisation, priming, and construction and use of schemas; and the analysis of metaphor from a conceptual standpoint (Ungerer and Schmid 1996).

Typology is the branch of theoretical linguistics concerned with the identification and analysis of patterns across different languages. Contemporary typology as a field was given its initial impetus by the work of Joseph Greenberg. Greenberg (1963) looked at a sample of thirty languages, assessing them on a range of features. He then interrogated this dataset to see what patterns emerged that might point to universal characteristics of language. Greenberg observed, for instance, a correspondence between the basic word order of a language, and the relative order of nouns and adpositions (that is, whether the language has [*preposition* + *noun*] phrases like English or [*noun* + *postposition*] phrases like Hindi). It turns out that languages with subject–object–verb order are overwhelmingly likely to have postpositions rather than prepositions. There is no a priori reason to expect this correspondence, and it therefore represents a valuable generalisation about language. Nowadays, typologists continue to make heavy use of Greenberg's basic technique – assemble a large sample of languages; gather information on their grammars; identify patterns or trends across languages – although there has of course been significant methodological development since the 1960s. For example, the universals that are identified are nearly always statistical tendencies rather than absolute patterns. Furthermore, finding explanations for cross-linguistic patterns is as important to contemporary typology as identifying the patterns themselves.

Finally, there are aspects of functional linguistics not directly classifiable as either cognitive or typological in nature, although they have links to both these forms of linguistics. For instance, an extensive theory (or set of theories) of

Functional Grammar has been developed (perhaps most notably by Dik 1997), which calls on the functions to which language is put – that is, features of semantics, pragmatics or discourse – as explanatory factors for observable features of grammar. In particular, it is often argued that the relationship between form and function in grammar is non-arbitrary, that is, that there exists a degree of *iconicity* in the relationship between the formal realisation of particular linguistic structures and the meanings the structures convey, or the functions those structures serve in communication. For example, Givón (1995: 125–6) points out that in English structures where one clause functions as the complement of a verb in another clause, there is a parallelism between the degree of clause integration (how much the two clauses behave as if they were a single complex clause: a formal feature) and the degree of event integration that is conveyed (how much the two events referred to by the two verbs are conceived of as a single complex event: a semantic feature). This can be seen, for instance, in the comparison between structures such as *she made him leave* and *she wished that he would leave*. The former is more structurally integrated than the latter, in that the independence of the second verb *as a clause* is much less, and refers to a semantically more integrated event (causation is perceived as more integrated than a preference).

Another important idea in functionalist theory, when applied to the diachronic study of language, is grammaticalisation – the process whereby over historical time, discourse-pragmatic patterns become increasingly fixed in structure, phonetically reduced and semantically generalised, ultimately becoming elements of grammar rather than discourse. We have already discussed in Chapter 5 how the techniques of corpus linguistics have been used to study grammaticalisation; for example, grammaticalisation is one of the explanatory factors called on by Leech (2004) in his account of recent change in English grammar. It is not a coincidence that grammaticalisation is both an important concern of functionalist theory and a focus of research in corpus linguistics. Although the tradition of English Corpus Linguistics which we reviewed in Chapter 4 is primarily a descriptive rather than a theoretical approach, it shares sufficient conceptual bases with functionalist theory – the rejection of Chomskyan approaches, the emphasis on language in use – that the subdisciplines inevitably coincide. This is sometimes identified explicitly; for instance Leech (2004b: 77–8) explicitly situates his diachronic analysis based on Brown, Frown, LOB and FLOB within a usage-based framework. Indeed, as we will argue later on, not just the corpus-based tradition of descriptive and historical English grammar, but in fact most if not all branches of corpus linguistics are coming together to an increasing degree with functionalist linguistics in the broad sense, in terms of research priorities, techniques and – most importantly – findings.

As we have noted, these various forms of functionalist linguistic theory are interrelated. For example, both functional and cognitive linguistics seek to explain linguistic form in terms of *meaning*, in the mind or in interaction. Likewise the cross-linguistic trends identified in typology are often explained by reference to functional or cognitive ideas – and evidence from typology can be important in

the formulation and justification of functional or cognitive theories. Accounts of Functional Grammar explicitly invoke typological and cognitive considerations (Dik 1997: 13–15; Givón 1995: 16–18) among their main bases. Likewise Croft's (2001) version of Construction Grammar may be fairly characterised as being both cognitive and typological in its underpinnings. While Croft seeks to explain language in terms of constructions, which are cognitive entities, much of the impetus for his theory, as well as much of the evidence he calls on to support it, comes from typology, such as the comparative analysis of different languages' inventories of part-of-speech categories.

We will see further examples of the pervasive interrelations between different approaches to functionalism in the following review of the impact corpus-based methods have had in these schools of theoretical linguistics.

7.3 Corpus-based research from a functionalist perspective

Much of the corpus-based research that we have discussed in earlier chapters could be described as 'functionalist' in the broadest sense, in terms of its theoretical stance. For example, Biber's multi-dimensional approach to text-type variation (see section 5.4) seeks functional explanations for formal (grammatical) differences – exactly the major concern of functionalism. Likewise, as we mentioned above, the role of grammaticalisation in much corpus-based research into the history of the English language (see section 5.2) makes such diachronic corpus linguistics likewise, in effect, a functionalist enterprise. However, in this section we will focus on research that has exploited corpus-based analyses within studies addressing core aspects of functionalist theory.

The exploitation of corpus data by theorists, especially syntacticians, working from a functionalist perspective has taken a number of forms. These vary greatly in the degree of similarity they show to the methodologies of corpus linguistics as a subdiscipline. For instance, before corpus linguistics emerged as a methodology, it was not uncommon for any set of linguistic examples – even one collected on an ad hoc or arbitrary basis – to be described as a 'corpus'. There are many examples of functionalist-oriented research where such a 'corpus' is used. For instance, Fox (1987) uses a 'corpus' of over a hundred relative clauses, drawn from an unspecified number of transcribed conversations, in a study of the noun phrase accessibility hierarchy. Downing (1993) uses around two hundred examples of two variant numeral constructions in Japanese, drawn from novels, oral narratives and transcribed conversations, in a study which demonstrates that these two constructions actually have different usage, in terms of the status of the information (new or given: see below) that they are used to express. Birner's (1994) study of non-interrogative subject–verb inversion in English (sentences with the order XVS . . . instead of SV . . .) is based on a 'corpus' of 1,778 examples of this structure which Birner and others happened to encounter and notice in various written and spoken sources.

However, in some studies using relatively small and arbitrary datasets described as 'corpora', particularly in the 1980s and 1990s, we may see the beginnings of a movement towards the incorporation of corpus data (in the usual sense) into functionalist-theoretical analysis. For example, Carden's (1982) study of 'backwards anaphora' (i.e. cataphora) demonstrates that it *is* grammatical – contrary to some scholars' earlier assertion – for pronouns to occur in contexts where the preceding discourse does not provide an antecedent for the pronoun. This demonstration is based on a 'corpus' of twelve texts (six children's books, a history book by Churchill, two adult fiction books, two newspapers and a book of Yeats); although this is too small and too arbitrary a selection of texts to be considered balanced and representative in the sense we discussed in Chapter 1, Carden's study does clearly show an awareness of the importance of representativeness. Carden also undertakes some (moderately rudimentary) quantitative analysis of the data. Siewierska's (1993) study of factors determining the word order in Polish transitive clauses, using data drawn from fourteen texts including fiction and non-fiction, likewise demonstrates a concern for sampling techniques across this relatively small corpus. Siewierska uses random sampling of clauses to manually collect sufficient examples of each of the six possible orders of subject, verb and object allowed in Polish. Siewierska's analysis of this data is founded on the quantification of functional factors across this set of examples: she 'scores' a range of factors that may affect the ordering of clause elements in Polish, including different measures of the size and *information status* of the subject and object. The information status of a noun phrase can be *given* or *new*. A noun phrase whose information is *given* refers to an entity which has already been brought into the discourse – the noun's meaning is active in the minds of the speaker and hearer. By contrast, a noun phrase which introduces into the discourse a previously unmentioned entity has *new* information. Siewierska's scoring methodology allows her to argue that information structure – that is, the tendency of a sentence's *topic* (usually given information) to precede its *comment* (usually new information) – is more important in determining transitive word order in Polish than is the *weight* of the subject and object – that is, the tendency of heavy (longer, more complex) NPs to occur later in a clause. Sun and Givón (1985) use a similarly extremely small text collection (twenty-five pages from a novel plus 55 minutes of transcribed speech) to quantify the effect of NP definiteness on object–verb versus verb–object ordering in Mandarin. On the basis of their statistical analysis, they are able to show that strong claims made in the prior literature on introspective grounds about this relationship – namely, that definite objects tend to precede the verb and indefinite objects tend to follow it (Li and Thompson 1975) – are in fact not true. Rather, the verb–object order is overwhelmingly most common in both speech and writing and, they argue, Mandarin is thus 'as rigid a [verb-object] language as, say English', with object–verb order functioning as a contrastive or emphatic device (Sun and Givón 1985: 344). The combination of qualitative (clause-analysis) and quantitative (statistical) methods in Siewierska and Sun and Givón's text-based studies anticipates the later use of

more extensively corpus-based methods in functionalist syntax, corpus analysis being inherently both qualitative and quantitative.

We may next consider studies that actually use large and/or standardised corpora of the kind we have discussed in earlier chapters. These may be divided on the basis of how comprehensively the corpus evidence is addressed. Some studies of syntax treat the corpus solely as a repository of examples, without taking a systematic approach to addressing the evidence of the corpus as a whole. For example, Declerck and Reed (2000) draw examples of English clauses marked by the conjunction *unless* from the COBUILD, Brown, LOB and ICE-GB corpora; Kaltenböck (2003) uses a very small number of examples also drawn from ICE-GB in a study of *it* in English; Toivanen (2002) analyses examples of the Swedish 'directed motion construction' found in the Swedish PAROLE[2] corpus; and König and Siemund's (2000) study of English reflexive pronouns draws on Brown and the BNC. In studies of this kind, the lack of systematicity inherent in the analysis of *only* a small selection of sentences means that the concern for 'total accountability' to the data that is typical of corpus linguistics (see section 1.6.1) is not in evidence.

Other research uses corpus data in a more systematic way. For instance, Valera (1998) examines the phenomenon of subject-oriented adverbs, that is, adverbs which structurally speaking modify a predicative adjective but which function to describe an attribute of the subject (e.g. *she was viciously unkind* implying *she was vicious and unkind*). Valera's dataset is drawn systematically and comprehensively from LOB, based on a concordance of all words tagged as adjectives that are modified by another word ending in *-ly*. He investigates, in a partially quantitative way, if any syntactic factors influence whether an adverb–adjective combination has a subject-oriented meaning or not, concluding that in fact none do, and that the lexical semantics of the words involved, and the compatibility of these meanings, is the major factor. As Valera's study illustrates, a typical concern of this type of research is tracking the influences of semantic, pragmatic or processing factors on syntax – that is, identifying *functional* motivations for grammatical form, the central focus of functionalist theory. The studies by Siewierska (1993) and Sun and Givón (1985) which we cited above are of this type (Siewierska looking at the impact of information structure and processing difficulty on word order, Sun and Givón looking at the impact of object definiteness on word order), although they work, as noted earlier, with very small datasets. Like Siewierska (1993), Arnold *et al*. (2000) investigate the relative importance of constituent weight and given/new information status on word order variation, looking at postverbal word order in English in this case. However, their study was able to use a large corpus (the Canadian Hansard corpus, a collection of transcribed parliamentary speeches), which they searched automatically, although further manual filtering of the results was needed. Their statistical analysis of the relationship between the newness and heaviness features and the word order of each search result shows that both constituent weight and givenness/newness correlate with the word order, and that each has a separate effect – that is, the fact

that both correlate is *not* simply due to the fact that new information tends to be expressed in heavy NPs and given information in light NPs.

The general approach that Arnold *et al*. employ can be summarised as follows: investigating the relationship between functional phenomena (such as semantics, givenness, processing difficulty) and formal phenomena (such as word order) by analysing these features across a set of instances resulting from a corpus search, and subjecting the outcome of this exercise to quantitative analysis, using the statistics of significance testing, correlation or regression. This approach has been used by many other studies. By way of non-exhaustive exemplification, we will briefly review McKoon and Macfarland (2000), Temperley (2003) and Hollmann (2005). McKoon and Macfarland's (2000) study takes a similar approach to Arnold *et al*. (2000) with regard to linking semantic and grammatical factors, although the focus is not on word order. McKoon and Macfarland use a combination of corpora amounting to 180 million words to look at the production of two semantic classes of verbs in English: internally caused change-of-state verbs (such as *bloom* or *flower*) and externally caused change-of-state verbs (such as *break*). They examine a number of different features for each example of the verbs they investigate, including the transitivity of the clause and the semantic type of the nouns that function as subjects and objects of the verb; they conclude that there is no difference in the transitivity of the two types of verbs, but that the ranges of semantic types of nouns that are the subject in transitive clauses is different between the internal-causation verbs and the external-causation verbs.

Temperley (2003) investigates a type of syntactic ambiguity in English relative clauses. This is the phenomenon where, if a zero relative pronoun is used, it may be possible for the first word of the relative clause to be interpreted as part of the main clause; Temperley gives the example phrase *the biological toll logging can take*, where the first four words are ambiguous on an initial reading – *logging* may be the head noun of the NP or the subject of the upcoming relative clause – the ambiguity only being resolved on the word *can*, which as a modal verb indicates that the word before it is more likely to have been a subject. Using the Penn Treebank (Marcus *et al*. 1993), Temperley examines all relative clauses where a zero pronoun is possible, extracted using a purpose-written computer program, and looks at a series of syntactic features of each instance, testing his hypothesis (that zero relative pronouns will be less common where the kind of temporary syntactic ambiguity in *the biological toll logging can take* would result) against a competing hypothesis, that the use of an anaphor as the relative clause subject motivates the use of zero relative pronouns. Temperley's statistical analysis shows that both these hypotheses are actually correct, in that both factors contribute independently to determining whether or not a zero relative pronoun is used; by contrast some other factors, such as the length of the relative clause, are found to have only non-significant effects on whether or not a zero relative pronoun is used. Finally, Hollmann (2005) looks at the active and passive versions of the English periphrastic causative, i.e. clauses of the form *make someone do something* and *be made to do something*, respectively. Hollmann searches the

BNC for examples of each, and then analyses a subset of the examples retrieved for a set of features related to functional or semantic qualities of the causative – including the type of causation, the punctuality of the aspect and the directness of the causation – applying a score for each parameter on each example. He finds significant differences between the active and passive causatives on some but not all of these parameters, thus allowing him to evaluate which semantic features are relevant, and which irrelevant, in a speaker's choice between the active and passive forms.

The studies we have cited so far in this section have mostly been undertaken by theoretical linguists seeking to incorporate corpus evidence into their analyses. The creation of links between corpus linguistics and functional theory has, of course, also occurred in the other direction, with corpus specialists using, and/or extending, functional theories in order to adequately characterise their data. A good example of this is the work on aspect in Mandarin Chinese undertaken by Xiao and McEnery. The earlier stages of this research (McEnery *et al*. 2003) took an approach to aspect in Mandarin very much typical of work studying variation in grammar across text-types in English datasets such as the Brown Corpus (see section 5.3.2). Unlike the languages of Western Europe, Mandarin lacks tense on verbs; but it has a wider range of aspect markers, which are collectively rather more frequent than the markers of perfect and progressive aspect in English – although the patterns of differing frequencies across genres are similar between Mandarin and English. However, a full account of aspect in Mandarin (and how it contrasts to English) required the extension of the analysis to involve matters of syntactic–semantic theory, leading to the development of what Xiao and McEnery (2004a, 2004b; see also Xiao 2008) call a 'two-level model of situation aspect'.

Situation aspect is concerned with the semantics of the time-structure of a particular state-of-affairs (a 'situation') that can be stated in language. Situation aspect is not, for the most part, a matter of grammatical aspect markers such as English *be* + *ing*-participle for the progressive aspect or *have* + *en*-participle for the perfect. Rather, situation aspect is concerned with the semantic features of the events being discussed, as speakers conceive of them. Previous theoretical work on situation aspect reviewed by Xiao and McEnery had proposed a range of different classifications of situation aspect according to various semantic features of the situations in question. For example, a distinction is commonly made between situations that are conceived of as having a duration in time (like *reading a book*) and situations that are conceived of as instantaneous (like *reaching a destination*). The latter type of situation is called an *achievement*; situations with a duration include *actions* (such as *reading a book*) and *states* (such as *standing still*). Xiao and McEnery expand prior models on the basis of their data by adding two factors not included in earlier work on situation aspect: whether the situation is bounded by an end-point in time, and whether the situation has a result. As a result they identify six different types of verb on the basis of the aspect of the situations those verbs describe, and eleven different types of situation that can be expressed at the sentence level by combinations of verbs and their arguments

and adverbials. This is thus a 'two-level' model because it classifies aspect at the lexical level (classes of verbs) as well as the sentence level.

Critically for our purposes, Xiao and McEnery's model of situation aspect is not formulated on the basis of, and supported by the evidence of, individual example sentences. Rather, they propose a series of *tests* for the presence of different semantic features of aspect in Mandarin and English – for instance, only verbs that have a duration are compatible with the Mandarin aspect marker *-zhe* (Xiao and McEnery 2004b: 334). These tests are evaluated quantitatively on the basis of exhaustive analyses of the relevant items in English and Mandarin corpus data. Furthermore, the interaction of different semantic features related to aspect is examined quantitatively in the corpus data, allowing rules to be devised describing how these features interact to create situation aspect at the sentence level (for example, the combination of different types of verb with count and non-count nouns – Xiao and McEnery 2004b: 334). In some cases this involves the manual analysis of relatively small amounts of corpus data. As a consequence of the corpus analysis, Xiao and McEnery claim to have created a more refined classification of aspect in verbs and sentences, one which, on the basis of its applicability in both English and Mandarin, is language-independent – although the tests used to identify different semantic features, such as compatibility of a verb with *-zhe*, are entirely language-specific (Xiao and McEnery 2004b: 360–1). In addition to this work on aspect in Mandarin, Xiao and McEnery have undertaken similar work on negation and passivisation in Mandarin based on the same corpus data (McEnery and Xiao 2005a; Xiao and McEnery 2008, 2010).

Xiao and McEnery's approach to testing different facets of their model of aspect against corpus data, in as exhaustive a manner as possible, is a variation on the procedure utilised by the functionalism-oriented studies cited earlier in this section: as comprehensive a set of examples as possible of some formal feature is extracted from a corpus; one or more semantic or other functional features is analysed for each example; and the results of this exercise are summarised and analysed on a quantitative basis. Thus, in all this research, we see that the joint quantitative–qualitative analysis typical of corpus linguistics lends itself very readily to the study of the functional-formal links, and the complex interactions among such links, sought within functional theory. From both sides, then, there is an impetus to bring together the practices of these two subdisciplines; and it is clear that this process is both fruitful and mutually enriching.

7.4 Corpora and typology

To a degree, it could be argued that all typology is inherently corpus-based. The raw data for typology is observations made about the grammar of a very wide collection of languages. These observations in turn derive from the work of field linguists. Field linguistics predates corpus linguistics but shares its

focus on empirically observed examples of language. It would be possible, in fact, to classify any field data that takes the form of a collection of sentences as a small 'corpus'. Admittedly, some field data is collected in ways which produce language that would be considered unnatural by the standards of corpus linguistics. Native-speaker informants in fieldwork have often been prompted to the degree that the resulting sentences must be considered elicited data, or even in some cases invented examples. Nonetheless, the methodologies of field linguistics and corpus linguistics coincide, although they have typically been undertaken as wholly separate enterprises. Moreover, some notions at the heart of typology are explicitly reliant on quantitative analysis of language usage; the most obvious case of this is the concept of Basic Word Order. The Basic Word Order of a language is the relative order of subject, object and verb in prototypical transitive clauses, often represented via abbreviations such as SOV or SVO (so we may say that English is an SVO language, Welsh a VSO language and so on). However, in many languages different word orders may occur – even English, which is relatively inflexible in terms of word order, allows OSV order in certain pragmatically marked contexts.[3] Therefore, Basic Word Order is defined as the *most frequent* word order in prototypical, transitive main clauses where both subject and object are full noun phrases (not pronouns). This kind of quantitative determination must obviously be made on the basis of relatively large amounts of authentic, naturally produced language data – that is to say, a corpus of some description. Basic Word Order is a primary parameter used when looking for patterns across different grammatical features in a sample of languages. For instance, in the example from Greenberg (1963) we cited above, the Basic Word Order SOV is linked to the presence in a language of postpositions rather than prepositions. So an effectively corpus-dependent notion is central to much typological analysis.

But the notion that corpus-like analyses are at the heart of typology cannot be upheld without substantial provisos. Typological analysis is typically based on what has been reported about the structure of the languages surveyed, rather than on direct analysis of collected data in those languages. Where a typological analysis does resort to the underlying field data, it is common for only a very few example sentences per language to be considered. This is unsurprising: for most languages, the available data is limited. While a very few languages, such as English, have an immense literature, a hugely greater number of languages are described in only one or a small handful of studies, or remain undescribed. It is thus, for instance, in practice impossible for a judgement about the Basic Word Order of a language to be actually based on quantitative corpus analysis in all cases. Mostly, it must be based on the reported judgement of the field linguist who produced the description of the language. So although typology is founded on empirical facts about languages as observed in the world – and is thus empirical to a degree at least comparable with corpus linguistics – the data used in typological reasoning is not generally 'corpus data' in a sense that corpus linguists would recognise.

There are exceptions, however. Some typological findings have been made based on the analysis of relatively large, relatively natural corpora. We will discuss one extended example in the remainder of this section: the discourse basis of ergativity (Du Bois 1987).

Ergativity (see Dixon 1979, 1994) is a pattern that has been observed in case marking and other grammatical phenomena across a very wide variety of languages. Put simply, it concerns the relationship between subjects in transitive and intransitive clauses, and direct objects. In *accusative* languages (such as English or German) the intransitive and transitive subjects are treated alike by the grammar, with the direct object treated differently; whereas in *ergative* languages (such as Basque) it is the intransitive subject and direct object that are treated alike grammatically. Du Bois (1987) proposes that ergative systems come into being because of the influence of certain patterns in discourse – patterns which exist not only in the languages that have ergative grammatical marking, but in others too. Du Bois' evidence is, however, drawn mainly from data from Sacapultec Maya, a language spoken in Guatemala, which has an ergative system of marking the bound pronouns that are affixed to each verb.

Du Bois' analysis is based on a dataset which he describes as a corpus, and which might fairly be considered as such from the perspective of corpus linguistics, although the data is not entirely spontaneous, as Du Bois' (1987: 811) report of its collection makes clear:

> I showed [a brief film without dialog] in Guatemala to a group of native speakers of Sacapultec; afterwards, each speaker was taken individually into a separate room to be interviewed by another Sacapultec [. . .] The interviewer explained that she (or he) had not seen the film, and asked the speaker to tell what had happened in it. The ensuing narration was tape-recorded, and later transcribed by myself and my Sacapultec assistants.

Du Bois identified, for each noun or pronoun in his small corpus, its grammatical role, animacy and information status. On the basis of this data, he argues for what he calls *Preferred Argument Structure*: a strong tendency in discourse for transitive subjects to be pronouns rather than full nouns, for transitive subjects to be given information, and for each clause to have only one full noun phrase representing new information. These discourse factors distinguish the transitive subject from both the intransitive subject and the direct object, and thus can explain the existence of grammatical systems that make the same distinction – that is, ergative systems.

Du Bois' work is firmly within the functional-typological tradition of linguistics. However, if we consider it from the perspective of corpus linguistics, it can be argued to be corpus-based in a number of senses beyond the obvious (i.e. his use of a small spoken corpus). His method involves total accountability to his corpus (see section 1.6.1), in that every noun phrase in his dataset contributed to the analysis. His analysis is both qualitative and quantitative (see section 1.1):

qualitative in the identification of the features of each noun phrase, and quantitative in the compilation of this information into overall frequency statistics of different kinds. Although the corpus data in Du Bois' study is not wholly natural, similar findings have been produced using data that is more natural, and thus more corpus-like. Genetti and Crain (2003), for instance, use a collection of ten spoken narratives in Nepali, which were collected without the uniform stimulus used by Du Bois, to observe broadly similar discourse patterns. Their texts include folk stories, ghost stories and personal accounts – close kin, in other words, to the genres of narrative texts found in English corpora such as LOB and the BNC, albeit spoken rather than written. So in terms of naturalness, Genetti and Crain's study represents a step closer to a corpus-linguistic style of data collection.

This example has shown how corpus-based analysis can in principle inform explanations of typological patterns. The fact that corpus linguistic methods have not, to date, had a major role to play in most typological investigations may be due, more than anything else, to issues of data availability. When corpora had only been constructed for a small number of languages, most of them in the Germanic or Romance families, it was difficult for typology – a field which rests above all else on comparing large numbers of languages from different families – to make use of corpus data. Nowadays, however, large corpora are becoming available for more and more languages. It would be possible, for instance, to repeat Genetti and Crain's study using some subset of the Nepali National Corpus (Yadava *et al*. 2008) – millions of words of entirely natural data.

7.5 Corpora and cognitive approaches to linguistics

Several different strands of work have brought together corpus methods and cognitive linguistic theory. One very productive branch of research has been the exploitation of corpus data within the field of Conceptual Metaphor Theory, which will be discussed in the next section. In this section, we will look at other studies that have used corpus data within a cognitive-linguistic framework for the analysis of syntax, semantics and the interface between them.

7.5.1 Cognitive approaches to syntax

Most corpus-based cognitive work on syntax has worked from the idea of a construction, as used in both Construction Grammar (Goldberg 1995; Croft 2001) and Cognitive Grammar (Langacker 2008). The idea of a grammatical *construction* is most readily explained by way of contrasting it to the idea of a grammatical *rule*. In formalist linguistics, the goal of syntax as a discipline is to identify the rules of the grammatical system – that is, the principles and constraints that determine which potential sentences in a given language are grammatical and which are not. An example would be the phrase-structure rules

used in early generative grammar (Chomsky 1957, 1965). These rules, such as S → NP VP ('a sentence consists of a noun phrase and a verb phrase') or PP → P NP ('a preposition phrase consists of a preposition followed by a noun phrase') can generate a certain set of syntactic forms but not others. Although the way that rules are expressed varies drastically in different formalist theories, one thing that is consistent is that rules have no meaning. This is due to the separation of semantics and syntax involved in many formalist theories. The syntactic rules simply generate a structure that defines an arrangement of grammatical classes; the *meaning* resides with the lexical items that are inserted into the 'slots' in the syntactic structure. The meaning of the overall sentence then arises from the interaction of the lexical items in the structure into which they have been inserted. This way of seeing grammar is not dissimilar in principle to an algebraic equation, where variables such as *x* and *y* can represent any number, and the nature of the equation is entirely independent of what numbers you might decide to slot into it. Unsurprisingly, many formalist models of syntax are markedly mathematical in their conceptual apparatus.

This separation between syntax and semantics – abstract rules with no meaning on the one hand, and meaningful words in the lexicon on the other – is not universally accepted. Cognitive linguists and other functionalists have long argued that grammatical structures *do* actually have meaning (see, for instance, Bolinger 1977). The term *construction* is used in place of the term *rule* to highlight this distinction, among other reasons. A construction is a unit of grammar that has a meaning, just like a word. It may contain some specific, concrete words along with grammatical slots into which other constructions or words can be inserted (such as the English construction *there BE (something)* indicating existence); or it may consist wholly of abstract slots (such as the English ditransitive construction *(someone) (verb) (someone) (something)*, which has a meaning related to transfer; e.g. *I give him a ball*). Knowledge of language consists of knowing an inventory of meaningful elements – some of them words, some of them abstract constructions. In the words of Tomasello (2003: 6), 'mature linguistic competence' consists of a 'structured inventory of constructions'. The basic ideas of Construction Grammar can be summarised as follows:

> a construction is any linguistic expression, no matter how concrete or abstract, that is directly associated with a particular meaning or function, and whose form or meaning cannot be compositionally derived. The linguistic system is then viewed as a continuum of successively more abstract constructions, from words to fully-fixed expressions to variable idioms to partially filled constructions to abstract constructions.
>
> (Stefanowitsch and Gries 2003: 212)

An approach to grammar based on constructions rather than rules fits well with the general precept of functionalism, that form and function (meaning) are linked. In fact the notion of the *construction* as the basic unit of grammar is often employed

by researchers in functional or cognitive linguistics not working directly on theories of Construction Grammar – including researchers working at the intersection of corpus linguistics and cognitive linguistics. Notably, the work of Stefan Gries and Anatol Stefanowitsch in combining the framework of Construction Grammar with quantitative analysis of corpus data has not only contributed to the understanding of how constructions behave in text, but also achieved the rare feat of making an important addition to the corpus-linguistic methodological toolbox.

The concept introduced by Stefanowitsch and Gries (2003) is the *collostruction* (the term is a blend of *collocation* and *construction*). A collostruction, like a collocation, is a co-occurrence link between two items in a corpus that occur significantly more frequently together than chance would predict. But while a collocation is between two words that occur frequently in proximity to each other, a collostruction is between a construction and a word that occurs frequently in one of its 'slots'. So for instance, the ditransitive construction has four slots: its subject, object and indirect object, plus of course the verb. Stefanowitsch and Gries (2003: 227–30) investigate which verbs tend to occur in the verb slot of the ditransitive, and find – perhaps not surprisingly – that the lemmata most significantly attracted to this slot include *give*, *tell*, *send*, *offer* and *show*, that is, the verbs often considered to form a category of ditransitive verbs in English. As a result, Stefanowitsch and Gries argue that ditransitivity is not actually a feature of these or any other verbs; rather, our perception of a class of ditransitive verbs is an epiphenomenon of their attraction to the ditransitive construction. Stefanowitsch and Gries introduce the term *collexeme* to describe a lemma significantly attracted to a particular construction. So, they describe *give*, *tell*, etc. as collexemes of the ditransitive construction. The converse relationship is described by the term *collostruct*: the ditransitive is a collostruct of *give*. The link between a collostruct and a collexeme is a collostruction.

The statistics used in calculating significant attraction between a construction-slot and the words that occur in it are different in detail, but not in major principle, from those used in calculating wordform collocations based on textual proximity (see sections 2.6.2 and 6.2). However, the methodology by which the base frequencies are collected for the calculation is rather different. Firstly, it is necessary to identify in a concordance all the instances of a given construction in the corpus. Constructions, however, can be difficult to search for. Even when a construction contains some specific words that can be queried – for instance, the *there-be* existential construction could be located by searching for *there* with an instance of *be* somewhere nearby – the searches are likely to be imprecise and require manual post-processing. A search for *there* with any form of *be* within three words each way, for instance, would catch actual *there-be* sentences such as *There might be trouble tonight* or *Is there anybody home?*, but it would also catch sentences such as *He's going there tomorrow* which is not an instance of *there-be*. So extensive manual analysis is needed, although if part-of-speech tags are available it may be possible to use them to assist the disambiguation. In the case of fully abstract constructions such as the ditransitive, it is even harder to

search reliably, even in a POS-tagged corpus. We could not, for instance, search for a NOUN VERB NOUN NOUN sequence and expect to find more than a tiny subset of the ditransitives in a corpus. This is because each of the slots can contain, as well as its 'main' word, a complex of other constructions. For example, the subject and object slots in the ditransitive can contain constructions associated with the noun phrase, such as relative clauses, premodifying adjectives and so on. Even when a concordance has been manually reduced to all and only the instances of the relevant construction, it is still necessary to identify, once again often manually, the correct word within the slot for the analysis. So for instance, in the sentence *He hasn't sent her anything yet*, the 'verb slot' of the ditransitive construction contains *hasn't sent*. But the word that 'counts' for the collostruction analysis is not *has* or *n't* but *sent* (and it is the lemma of *sent*, i.e. *send*, that is counted).

Stefanowitsch and Gries deal with the various problems of searching for constructions and the words in their slots in two ways: firstly, as noted, by extensive manual analysis of concordances, and secondly by using manually annotated corpora with full syntactic parsing. In a subsequent, extended analysis of the ditransitive construction, Gries and Stefanowitsch (2004) use the parsed ICE-GB corpus to identify instances of this construction and others they are interested in. This is probably the main difference between collostruction and traditional statistical approaches to collocation. The calculation of collocation statistics, relying only on proximity, can be entirely automated. The calculation of collostructions, relying on the identification of constructions and their slots, can be partially automated but it seems likely that some degree of manual analysis will always be necessary. This means that it is not likely that standard corpus analysis software (see Chapter 2) will ever possess a button that says 'press here for collostructions'. However it *is* possible, and in our view desirable, that corpus tools will come to integrate mechanisms for carrying out the automatable steps of a collostruction analysis. Of the current widely used corpus tools, SketchEngine (see section 2.5.4) is the one that comes closest to integrating the collostruction procedure.

Collostruction, as defined by Stefanowitsch and Gries, is not an entirely new idea. Some work described by its authors as concerning *collocation* pre-empted some of the ideas behind *collostruction*, at least in part. In Cognitive Grammar, for instance, the term *collocation* is used (e.g. by Langacker 2008: 20) to describe an instantiation of an abstract construction that is itself an established unit in the language. So if *(verb) (someone) in the (body-part)* is a construction, the expression of the form *poke (someone) in the eye* is a collocation. In Stefanowitsch and Gries' terms, this could perhaps best be described as two collexemes jointly associated with a single construction. In this case, then, we could perhaps see collostruction analysis as systematising and implementing existing ideas in cognitive linguistics. Furthermore, although most work on collocation within corpus linguistics has been based solely on proximity, there have been exceptions to this. As we noted in section 6.2, outside the neo-Firthian tradition there has

been work where a collocation is defined as existing not simply between two words that are *near* to one another in the text, but rather between two words that stand in a particular grammatical relationship to one another: for example a premodifying adjective and its head noun, a verb and its object, or a verb and an associated preposition phrase. Evert (2008) refers to this as collocation using 'syntactic cooccurrence' as opposed to the 'surface cooccurrence' of the traditional, proximity-based approach to collocation (see Krenn and Evert 2001 for an example). This syntax-based approach to collocation also approximates a collostructional analysis; the main difference is the absence of the theoretical commitment to some variety of Construction Grammar. It is that theoretical commitment which leads to the possibility, in a collostruction analysis, of calculating associations between words and slots in *abstract* constructions, which is surely the most exciting possibility opened up by this form of corpus analysis.

Many of Stefanowitsch and Gries' papers on collostruction have used the ditransitive as an example. This is perhaps not surprising. The English ditransitive construction is a frequent focus of interest in functional or cognitive studies, perhaps in part because of the need to identify a semantic or functional explanation for the variation between ditransitive and the *to*-dative within the framework where constructions are lexical items with individual semantic content, a need which Gries and Stefanowitsch (2004) attempt to address. Some of this work has indeed been corpus-based within a functionalist theoretical framework (e.g. Hollmann 2007; Siewierska and Hollmann 2007). The ditransitive has also been a key working example for other work bringing corpus methods to cognitive linguistics, for instance Mukherjee (2004), who uses the lexicogrammatical patterns around *give* in ICE-GB to argue that frequency, distributional patterns and facts of routine usage as observed by corpus methods should be part of the Cognitive Grammar model. In later work, Mukherjee goes on to exploit patterns of usage of the ditransitive as a means of contrasting varieties of English (Hoffmann and Mukherjee 2007), utilising the collostructional approach within this research (Mukherjee and Gries 2009). This reflects more recent work by Stefanowitsch and Gries, which has also focused on the use of collostruction as a method for analysis of dialectal (and other varietal) differences (e.g. Wulff *et al*. 2007; Stefanowitsch and Gries 2008).

7.5.2 From syntax to semantics

In the field of (lexical) semantics, which is another major concern of cognitive linguistics, a wide variety of corpus-based approaches has also been applied in recent years. A number of different strands of such work may be identified. Some research has been undertaken to problematise work in one theory or other of functional-cognitive semantics. For instance, Wierzbicka (1972, 1980) proposes an approach to semantics dubbed the 'Natural Semantic Metalanguage'. In this approach, complex word meanings are expressed in terms of simpler units of meaning, until a set of irreducible semantic 'primes' or 'primitives' are

arrived at – that is, units of meaning that cannot be explained or paraphrased using combinations of even simpler units of meaning. These semantic primes are considered to be aspects of conceptualisation and are argued to be language-independent – that is, they may be seen as universal aspects of human cognition. Kitis (2009) uses corpus evidence from English and Greek to argue against Wierzbicka's analysis of one particular sort of verb, namely verbs meaning *fear* or *afraid*. Kitis argues that in both languages, the verb *fear* (*fovume* in Greek) evolved over time, from an original meaning of 'to put to flight' or 'to scare away', to its contemporary meaning of 'to experience the emotion of fear'. This meaning shift is, according to Kitis, an example of metonymy, where the physical motion is metonymic for the emotion. This development, Kitis posits, is problematic for Wierzbicka's approach, where a meaning such as *fear* would be explained in terms of simpler semantic primes such as *feel*, *think* and *bad*: such a decomposition misses the phenomenon, evident in the historical corpus examples, of these meanings actually emerging from meanings related to physical responses and actions (i.e. running away, being put to flight).

In other cases, corpus-based functional-cognitive semantics has extended rather than problematised earlier work. For instance, Langacker's approach to linking syntax and semantics in Cognitive Grammar has formed the foundation of studies such as that of Kaleta (2009), who looks at the near-equivalent structures *begin to VERB* and *begin VERBing* in English. Using a random set of 250 examples of *begin* from the BNC, Kaleta builds on Langacker's schema for the meaning of *begin*, elaborating it and differentiating it for each of these two complementation patterns. The result of this analysis – that *begin to VERB* and *begin VERBing* differ in the type of 'boundedness' that they imply – is clearly in keeping with the general finding of functional-cognitive linguistics, repeatedly supported by evidence from corpus studies, that form and function are linked, and that syntax is meaningful.

One interesting, recently developed approach which likewise underlines the integration of syntax and semantics is the analysis of lexical semantics in terms of *behavioural profiles*, a notion introduced in its current sense by Divjak and Gries (2006), Divjak (2006) and Gries (2006b) as an expansion of Hanks' (1996) use of the same term for a similar but more restricted analysis. A behavioural profile is a corpus-derived characterisation of a word which allows, in particular, a close examination of meaning differences between pairs or sets of near-synonyms, or between distinct senses of a single word. The method by which behavioural profiles are built (Gries and Divjak 2009: 61–4) is conceptually simple, although details of the implementation can be rather involved. It involves getting a concordance of the word(s) under investigation, and annotating each concordance line for a series of morphological, semantic or syntactic features (these features are dubbed 'ID tags'), so that each instance of the word becomes a data point consisting of, in effect, a series of attribute–value pairs (such as tense = past/present/future, aspect = imperfective/perfective, semantic type of subject = concrete/abstract; and so on: see Gries and Divjak 2009: 62). For each feature, the frequencies of the different values are summed and calculated as proportions – for

example, the fraction of the total number of concordance lines that has past tense as opposed to present and future. These frequencies are then used as the starting point for a statistical analysis of differences between near-synonyms, or between different senses of a single word (where the figures for each sense of the word are calculated separately). A variety of statistical analyses, both basic and more sophisticated, may be used to investigate the data. One that Gries and Divjak normally apply is cluster analysis (Gries and Divjak 2009: 65–7). Cluster analysis groups together the words under investigation (or the senses of one word) in terms of their relative similarity, where the statistical measure of 'similarity' incorporates *all* the information from *all* the features that have been annotated on the concordance lines (see section 2.6.2 for a brief overview). Gries (2006b) does a cluster analysis of the several dozen different senses he identifies for the English verb *run* (for a highly common verb to have such a large number of meanings is not especially surprising); this analysis groups together senses that are similar at various levels in terms of their linguistic behaviours, which appears also to correspond to some intuitively salient distinctions among senses of *run*. For example, all the meanings of *run* having to do with motion (as in *to run off* or *to run away*) are grouped together; the meanings having to do with 'managing something' (as in *to run a company*) are grouped together; and the cluster analysis also distinguishes senses used transitively from senses used intransitively (Gries 2006b: 81–2). Working on the other issue, near-synonymy of different words, Divjak (2006) examines four different Russian verbs with meanings similar to 'to intend (to do something)'. Using a corpus of twentieth-century Russian literature, Divjak constructs a behavioural profile from a concordance of each verb, and then clusters the resulting data, to establish a 'family tree' of behavioural similarity among the verbs, showing which are most like one another and which are outliers. Divjak argues (2006: 46) that this is a valid alternative method for differentiating near-synonyms to traditional approaches reliant on introspecting about the words' meanings. Subsequent studies using this behavioural profiles methodology have looked at the polysemy of the verb *get* in English (Berez and Gries 2009), near-synonymy of verbs meaning 'try' in Russian (Divjak and Gries 2006) and near-synonymy and antonymy in English size adjectives (such as *big*, *little*, *great* and *small*: Gries and Otani 2010).

Studies such as these show how corpus data can be used to inform functional-cognitive approaches to lexical semantics. However, we have not yet touched on one of the most revolutionary aspects of the cognitive approach to semantics, namely the analysis of metaphor in conceptual terms. This area of research, and the role that corpus data may play in it, are the topic of the following section.

7.6 Corpora in the analysis of metaphor

As noted above, another important facet of cognitive linguistics is Conceptual Metaphor Theory. Traditional approaches to metaphor as a linguistic

phenomenon focused mostly on the creative use of metaphor in (primarily) literary texts; that is, metaphor was treated mostly as a stylistic device. Scholars had long been aware of highly conventionalised or 'dead' metaphors – metaphorical expressions so commonly and habitually used that they no longer really attract attention as metaphorical, although they clearly cannot be interpreted literally; for example *he's bursting with pride*, or *she's buried in paperwork* (see Ungerer and Schmid 1996: 116–18 for an overview). Conceptual Metaphor Theory (CMT), proposed by Lakoff and Johnson (1980) among others, inverts the traditional approach to metaphor by focusing mainly on this everyday, conventionalised use of metaphor. CMT is driven by the observation that the 'dead' metaphors are absolutely pervasive in everyday usage. Rather than a creative literary device, it is argued, metaphor is a part of the way we think, which is then reflected in the way we speak. For example, we often use our conceptualisation of concrete things to structure the way we think about abstract things. So part of our cognitive 'toolkit' is a group of metaphorical mappings such as TIME IS MONEY or LIFE IS A JOURNEY (by convention in CMT, statements of the form X IS Y in small capitals indicate that we use a source domain *Y* to help us think about a target domain *X*). Conceptual metaphors like these are realised in a range of metaphorical expressions, such as *I spent some time in France this year*, *we're wasting time here*, *I'm at a crossroads*, *she's found her true path in life* and so on.

As with Cognitive Grammar, at its inception much work in CMT was based on intuition and the analysis of invented examples (such as those given in the previous paragraph). It was therefore possible to criticise CMT as lacking empirical support from language usage data (see Deignan 2005: 27). However, in more recent years a great deal of work has been done to bring the framework of CMT together with the analysis of real language usage data, in many cases utilising corpus methods. At the most basic level, this could involve, for instance, replacing fabricated examples with examples from a corpus when illustrating a particular conceptual metaphor. For example, we could search a concordance of *life* for instances of expressions where it is conceptualised as a journey, rather than making up examples. With a sufficiently powerful concordancer, more specific searches can be used to search for particular types of likely-to-be-metaphorical phraseology. In the BNC, in a concordance of instances of *path* and *life* in the same sentence, we find examples such as the following, which vary from the fairly creative to the extremely conventional:

What do we talk about as we walk the path of life?
. . . as if he were a particularly prickly thorn in the path of her life.
. . . which he sees as the third and highest stage along the path of life . . .
. . . Thomas speculated in fictional form upon another path his life might have taken . . .
. . . difficulties of later life, as to discover how people find a path through them . . .
They've never crossed your virtuous path through life . . .
He . . . will follow the path of karate-do for life.

A discussion of the LIFE IS A JOURNEY metaphor cannot help but be enriched by the use of such examples. However, it is possible to go much further in the use of corpus data within CMT, enabling analyses that would be impossible without corpus data, and in some cases problematising the theory itself. A major proponent of this approach is Deignan (1999a, 1999b, 2005).

Deignan (2005) examines a range of phenomena associated with metaphor where investigating relevant corpus data reveals patterns of use that are not straightforwardly predictable from a CMT account of the source domains, target domains and expressions involved. Like much of the research in cognitive syntax and semantics discussed above, this involves extensive manual analysis of concordances of potentially relevant words or phrases; direct automatic retrieval of metaphorical language is currently not possible, and current corpus tools can only assist the procedure. One basic example of Deignan's approach is her investigation of the metaphor AN ELECTION IS A HORSE RACE (Deignan 2005: 27–31). This metaphor is realised by expressions such as *in the running*, *neck and neck*, *first past the post* and so on. What Deignan observes in the corpus is that there is no simple or regular transference of language from the source domain to the target domain. Rather, there are arbitrary conventions on how the conceptual metaphor is employed. For example, the expression *first past the post* has a specific meaning when used metaphorically (it designates an electoral system where the candidate with the most votes is victorious regardless of how many votes were cast in total). This meaning cannot be inferred from the literal meaning of the expression plus the AN ELECTION IS A HORSE RACE metaphor. Nor can the notion of being *first past the post* be used metaphorically in any other way than as a label for that type of electoral system.

Deignan makes similar observations with regard to the interaction of metaphor with grammar in corpus data. Metaphorical expressions are often grammatically distinct from the corresponding literal expressions in terms of the categories and structures they may occur in. For instance, the distinction between literal and metaphorical usage of a particular word may be linked to a part-of-speech distinction. Metaphors based on the source domain of animals may be realised by expressions where the animal-word has undergone conversion from a noun to a verb or adjective (Deignan 2005: 152–5). While it is easy to think of examples of this kind of metaphorical expression – for example, any use of *dog* or *badger* as a verb that we might come up with via intuition is likely to be metaphorical – what the corpus evidence shows is the quantitative predominance of the literal sense when the animal-word functions as a noun, and of the metaphorical sense when the animal-word functions as a verb or adjective. Such quantitative facts are inaccessible to intuition, but more interestingly are not, Deignan argues, predictable solely from the precepts of CMT. A similar pattern emerges from Deignan's survey of the interaction of metaphor with collocation (2005: 198–210). For example, Deignan looks at the word *price*, whose source domain is money but which is often used metaphorically, and finds that some of its collocates (e.g. *high* as a premodifier) co-occur with both the literal and the metaphorical senses of price. By contrast, some other collocates (e.g. *heavy* as a premodifier) are

restricted to occurring with its metaphorical sense, and others are restricted to the literal sense (e.g. *sale*). But overall, what is most remarkable is Deignan's finding (2005: 208) that almost all instances of *price* co-occur with at least one collocate or colligate which is unique to one or the other sense, and which thus disambiguates its meaning in context. Thus, in sum, Deignan establishes that metaphorical and literal meanings of any particular word are distinguishable by their co-occurrence with particular lexical and/or grammatical patterns. This may be seen as a special case of the argument that, in actual usage as observed in a corpus, words are never semantically ambiguous because linguistic context always disambiguates them (see Teubert and Čermáková 2004: 16–17); this motivates the neo-Firthian view of the unit of meaning, or lexical unit (see section 6.2) as the shortest stretch of words which is unambiguous as to meaning, which as we have seen is an important idea in the framework of Sinclair's Idiom Principle.

In these cases, and others, Deignan's repeated finding is that CMT is necessary, but not sufficient, for an explanation of the behaviour of metaphorical language as observed by corpus methods: '[w]hile corpus data are rarely, if ever, inconsistent with Conceptual Metaphor Theory, many of the detailed linguistic features of metaphor [. . .] are not easily explained by it' (Deignan 2005: 223). Other explanatory factors must be invoked to develop a full account of the observed usage patterns – whether ideas relating to extended lexical items, as above, or other cognitive or communicative theories (Deignan 2005: 164–6, 223). Thus, the status of CMT as an all-encompassing account of metaphorical language is problematised by Deignan's analyses. This exemplifies the invaluable role of the corpus methodology for testing theories about language against attested language usage.

7.7 Summary

In the 1980s and to a lesser extent the 1990s, it would have been largely fair to describe corpus linguistics and functionalist linguistics (broadly defined) as separate enterprises, divided by the data they were based on. Functional linguistics was largely based on the analysis of relatively small numbers of authentic examples, or, in some cases, fabricated examples; and on intuitions and grammaticality or acceptability judgements. This was sharply different from the comparatively very large textual resources being analysed by corpus linguists even in the 1980s. This division is, to a large extent, breaking down, as the research we have discussed in this chapter tends to illustrate. Functional-cognitive-typological research is more and more open to the use of larger and larger datasets; and, as this happens, the tools and techniques of corpus linguistics become more and more essential and integral to the analyses that are undertaken.

In the following chapter, we will review the use of corpus data in experimental psycholinguistics, which has similarly grown greatly over the past twenty years.

We will also show that there has been a convergence between the findings of the functionalist enterprise (and of psycholinguistics) and the main findings of the neo-Firthian school of corpus linguistics.

Further reading

Our sketches of the fields of functional theory, typology and various strands of cognitive linguistics have necessarily been very brief. If you are interested in finding out more about these areas, we would recommend going in the first instance to some of the excellent introductory texts that exist for these areas of linguistics. For typology, Whaley (1997) gives a brief and readable overview (which includes a discussion of the phenomenon of ergativity). For an introductory-level overview of the main concerns and primary explanatory ideas of cognitive linguistics, Ungerer and Schmid (1996) is an excellent read. For more on Conceptual Metaphor Theory, see Lakoff (1987). More advanced accounts of Functional Grammar are given by Givón (1995) and Dik (1997); for Cognitive Grammar, see Langacker (2008) and for Construction Grammar, see Goldberg (1995) or Croft (2001).

We are not aware of any detailed treatment in the secondary literature of, specifically, the role of corpora in functional or typological linguistics, so we can do no more than refer to the original research papers we have cited in the earlier part of this chapter, not all of which are accessible reads for non-specialists. From the perspective of a reader with more knowledge of corpus methods than functional theory, Xiao and McEnery's (2010) comparison of English and Mandarin Chinese is possibly one of the more accessible examples of research into the intersection of these fields. An overview of the convergence between corpus and cognitive linguistics is given by Gries (2006a). On collostructional analysis, the very first paper published (Stefanowitsch and Gries 2003) is probably the easiest read, although if you are interested in the full scope of the method, it is worth also looking at Gries and Stefanowitsch (2004) and Stefanowitsch and Gries (2005); for a summary of behavioural profiles, see Gries and Divjak (2009). The best source for corpus-based work in the area of Conceptual Metaphor Theory is Deignan (2005), reviewed extensively in this chapter; but see also Semino (2008).

Practical activities

(A7-1) In typology the Basic Word Order (BWO) of a language is a key feature for classifying that language's grammar. Ideally this should be established on the basis of corpus data. BWO is determined on the basis of *main*, *positive*, *declarative*, *prototypical transitive* clauses, where the subject and object are *both full noun phrases* (i.e. not pronouns or clauses) and the *object is definite* – clauses like *The architect built the house*, to give a fabricated example.

- How can you search for this *very specific* sort of transitive clause in a POS-tagged corpus? *Hint*: you will probably need to start with a less specific initial search, then narrow it down by manual filtering of the concordance result.
- If you have access to a parsed corpus, see if you can find out how to use parse-structure searches to locate these clauses – whether or not this is possible will depend ultimately on both the nature of the parsing system and your search tool.
- Try searching manually through a small sample of short texts (for example, two or three newspaper articles) for clauses that 'count' for BWO. How many did you find? Do you think this kind of manual search is more or less effective than the concordance-then-manual-filter method?

(A7-2) In section 7.5.1, we mentioned the *(verb) (someone) in the (body-part)* construction, as in, for instance, *Elizabeth poked Mrs Smith in the eye*. Work out how to run a search for instances of this construction. Note that this is probably best done with at least a moderately large, POS-tagged corpus! Once you have got a concordance of this construction, do a basic collostruction-style analysis of the words it co-occurs with.

- What kinds of lemmata appear in the *verb* slot? What kinds appear in the *(body-part)* slot? If you feel confident with the statistics, you can test the significance of these associations using the method of Stefanowitsch and Gries (2003) – if you don't know how to do a Fisher Exact Test, chi-square or log-likelihood will also work for the purposes of demonstrating the method. If you are not quite so confident, stick to frequency counts of the different collexemes.
- What relationships can you find between the verb that is chosen for the first slot and the body part chosen for the third slot? We noted that we would expect *poke* and *eye* to be linked, but are there other pairs? Are all the pairs you find things you could have guessed in advance, or are there surprises? Are the pairs explicable by semantics or world-knowledge, or are they arbitrary facts of phraseology?

(A7-3) It has often been observed that the English word *literally* can function as a marker of a metaphor – that is, the thing that is said to be *literally* happening isn't really happening at all; rather, it is a metaphor for something else that actually is happening.

- Run a query for *literally* in a general corpus. Analyse the concordance lines – in how many is *literally* used literally? In how many is it used metaphorically?
- Does *literally* tend to co-occur more with the creative metaphors of interest in traditional approaches, or with the conventionalised or 'dead' metaphors that are more of interest to Conceptual Metaphor Theory?
- Could we realistically make use of searches for the word *literally* to locate metaphorical usage in our corpus without specifying the metaphorical word itself in advance? Can you think of other possible marks of metaphor that might be used in this way?

Questions for discussion

(Q7-1) Functional and typological approaches to grammar are very strongly based in the analysis of the grammars of many different languages, from many different families. However, the languages for which very large corpora are available are (a) few in number and (b) drawn from only a fraction of the existing language families – Indo-European languages in particular being massively over-represented in the available corpus data. What potential problems does this raise for corpus-based functional-typological analysis?

- Is there any value for typology in contrastive studies based on fundamentally similar languages? (See, for instance, Hardie and Mudraya 2009, which compares adpositions in Nepali, English and Russian – all Indo-European languages.) How about contrastive studies that compare only a single pair of languages, like the English/Chinese comparisons of McEnery and Xiao – are these potentially useful for typology, or not?
- How feasible would it be to collect corpora in *all* the languages typologists are interested in? What are the pros and cons of focusing our effort here, rather than on analysing the corpora that we *do* have?

(Q7-2) In Chapter 6, we mentioned that the notion of *discourse* is defined in contrasting ways in Critical Discourse Analysis (CDA) and functionalist theory – reread section 6.3 if this is not fresh in your mind! Now that we have discussed functionalist theory, and seen some examples of how it uses the notion of discourse (for example as in Du Bois' 'discourse basis of ergativity'), we can pose the following question: are corpus methods a better match for one of these ways of approaching *discourse* than they are for the other? Why, or why not?

(Q7-3) One main finding of Deignan's analysis of metaphor based on corpus methods is that CMT is necessary but not sufficient as an explanation of the observed patterns of metaphorical language in the corpus.

- Is it a problem for CMT that much of human metaphorical behaviour, seen on the large scale, is not explained solely by the processes that CMT theorises? That is, if we have to bring in other theories to get a complete explanation, does that mean there is something *wrong* with CMT?
- Can the following two ideas be reconciled? (i) it is often the case that the metaphorical usage of a particular word is limited to specific phraseological contexts that are presumably learnt as idiomatic chunks and not analysed by speakers; (ii) conceptual mappings between a source domain and target domain occur at the conceptual level, not the linguistic level, and are productive, leading to many different metaphorical expressions.
- Is CMT compatible with the neo-Firthian theory that is used to explain the context-dependent nature of metaphor – especially the stronger versions of that theory, as espoused by Louw and Teubert, for instance?

8 The convergence of corpus linguistics, psycholinguistics and functionalist linguistics

8.1 Introduction

As we have seen in Chapter 7, functionalist linguistics in the broad sense (including cognitive linguistics) is increasingly making use of corpus-based methods, and in turn informing the analyses of corpus linguists. In this chapter, we will show that this phenomenon extends as well to experimental psycholinguistics. We will also discuss the implications of the rapprochement of functionalist linguistics and psycholinguistics with corpus linguistics with regard to the neo-Firthian school of thought which we surveyed in Chapter 6; we will argue that in the neo-Firthian school, this rapprochement with functional linguistics has taken a very different form. As we saw in Chapter 6, one of the bases of the neo-Firthian or so-called 'corpus-driven' approach is a rejection of non-corpus-derived theoretical frameworks. To explicitly adopt a functionalist theory as the basis for a corpus-driven study would be distinctly peculiar from the neo-Firthian perspective. Indeed, some of the stronger forms of the neo-Firthian position – such as that espoused by Teubert, for instance – explicitly reject the notion of a convergence of neo-Firthian corpus linguistics and functional or cognitive linguistics, with Teubert (2005: 2) claiming that corpus linguistics 'offers a perspective on language that sets it apart from received views or the views of cognitive linguistics, both relying heavily on categories gained from introspection rather than from the data itself'. Nevertheless, we wish to argue that such a convergence is in fact taking place, stemming on the neo-Firthian side from work by Sinclair and others from the 1990s onwards. Our basis for making this case is that, when we closely examine the findings of the most extensively developed neo-Firthian theories – in particular, Pattern Grammar and Lexical Priming – we will find that many of these conclusions have *also* been arrived at by one or more branches of functional linguistics or psycholinguistics. These congruent conclusions stem from wildly different sets of evidence and are, of course, expressed using very different descriptive apparatus. But certain fundamental insights – namely, the inseparability of lexis and grammar, and the nature of grammar as secondary to, and emergent from, lexis – have been arrived at by both functional linguists and neo-Firthian corpus linguists, largely independently of one another.

In this chapter, then, we have two main topics. Firstly, in section 8.2 we will consider the role of corpora in experimental psycholinguistics, as we considered their role in functionalism in Chapter 7. Psycholinguistics as a discipline is methodologically rather different to functionalist theoretical linguistics, but it shows signs of a similar trend with regard to corpus methods – that is, that over recent years there has been more and more use of corpus data within psycholinguistic research, and a convergence or rapprochement between the findings of psycholinguistic experiments and of corpus investigations.

Secondly, section 8.3 discuss the convergence of findings, regarding in particular the ontological status of grammar, lexis and language itself, between neo-Firthian corpus linguistics, functional linguistics and psycholinguistics.

8.2 Corpus methods and psycholinguistics

Overlapping cognitive linguistics (which we discussed in the previous chapter), but in many ways distinct from it, is the field of psycholinguistics – and in particular that branch of psycholinguistics whose methodology is mainly experimental. In the latter approach, the primary source of data is various types of laboratory tests on human subjects (or, as we will see later, computer models). While experimental psycholinguistics is not usually considered a branch of functional-cognitive linguistics, its fundamental methodological assumption – that the nature of language in the brain or mind can be investigated in much the same way that experimental psychology in general looks at other aspects of the nature of thought – is in accordance with the general tenet of functionalism that there is no absolute divide between form and function, between language and non-linguistic cognition.

Psycholinguistics is a very broad field, and there is absolutely no room here for a full review of it – nor even to treat comprehensively all research which has linked psycholinguistics with corpus data and methods. We must therefore confine ourselves to an extremely brief and purely indicative survey. To characterise psycholinguistics in very broad terms, we might say that it is focused on two primary issues (which are closely interrelated, as Ellis 2002 illustrates): language learning and language processing. There are other topics of interest of course, such as the evolution of the language faculty. However, we will limit ourselves here to looking at how corpora have been used in some psycholinguistic investigations into first language acquisition, second language acquisition and language processing.

8.2.1 Corpus data in experiments on language processing

Language processing has been investigated experimentally in a number of ways. Two that are reasonably common are *self-paced reading* experiments

and *eye-tracking* experiments. Both are means of investigating the speed with which particular segments of language are processed. In a self-paced reading experiment, participants work at a computer running a specially designed program. The computer shows one word of a sentence at a time to the participant, who presses a button to get the next word once they have read the word currently on screen. The program records the time for each button-press, so that the relative speed of reading for each word is known. Typically, after each sentence participants have to answer a (very easy) question about the content of the sentence – this prevents participants from just clicking through sentences without actually reading for meaning. The results of such an experiment can be used to infer what elements (morphological, syntactic or semantic) are processed easily, and which are more difficult and thus require more processing time. This in turn can give indications about what is actually happening in the brain. Although useful, self-paced reading experiments may potentially be misleading in that fluent readers do not typically read one word at a time, in sequence, without ever going back in the text. In fact, it is known that a reader's journey through a sentence of printed text can be quite complex, with multiple movements back and forth. This type of evidence is gathered in eye-tracking experiments (see Rayner 1998 for a review). Again, participants are given the task of reading sentences presented on a computer screen, but this time an entire sentence is presented at one time, and specialised video equipment records the movements of one of the participant's eyes as it looks at different positions in the sentence immediately after the sentence appears on screen. The resulting data is much richer, but correspondingly rather more difficult to interpret, than self-paced reading data.

These kinds of experiments may seem remote from the concerns of corpus linguistics. However, there are at least two ways in which corpus data can play an important role in the design and interpretation of such experiments. Firstly, corpus data can be used as a check on the naturalness of the language task that the experiment sets its participants. For instance, Frisson and Pickering (2001) summarise the results of a series of eye-tracking experiments aimed at investigating the processing of words which are ambiguous between a literal and a metaphorical meaning, when the part of the sentence prior to the ambiguous word does not provide sufficient cues to indicate which meaning is intended. But Deignan (2005: 114–17), in a review of this study, points out that in fact, such cases almost never occur in corpora of real usage: in *all* the examples she looks at, some aspect of the preceding context – possibly in an earlier sentence – indicates which meaning is intended. So, for instance, the word *campaign* literally relates to warfare and metaphorically relates to politics. In any given real example of *campaign* from a corpus, the prior context is overwhelmingly likely to give some indication whether a military campaign or a political campaign is intended; so by the time the reader gets up to *campaign*, it is already effectively disambiguated. On this basis, Deignan argues that if an experiment presents participants with a word such as *campaign* without any indication in the foregoing text as to whether it is literal or metaphorical, as Frisson and Pickering's experiment did, then that

experiment is actually 'forcing participants to tackle problems that are not faced in normal discourse' (Deignan 2005: 117). If this is the case, then it may be argued that while such an experiment may indeed tell us something interesting about the processing of ambiguously metaphorical words, it cannot tell us about the normal processing of language in use. We can see, then, that a corpus-derived awareness of how words (and other linguistic items) are *actually used* can serve as a useful anchoring-point for psycholinguistic experimentation. This is not to say that unnatural language should never be used in an experiment – there are cases where non-idiomatic language may itself be the object of study, for instance Millar's (2011) study of how errors in collocation, of the type made by non-native speakers of English, can affect processing speed in self-paced reading. What is undesirable is a situation where experimental tasks include highly unnatural language *without the experimenter being aware that this is the case*.

Secondly, corpus data can be used as a source of *frequency data* in the construction of test sentences in self-paced reading or eye-tracking experiments. Often, the test sentences used will not be drawn directly from corpus data, because the analysis of the resulting data may require certain aspects of the sentences to be controlled across different examples. For instance, if we are primarily interested in the time taken to process (say) the verb in a sentence, then we might well wish to control the length and syntactic structure of the preverbal elements (as well as, potentially, that of the rest of the sentence). We are unlikely to find such controlled sentences in a corpus! But even when invented example sentences are used, it is entirely possible for the creation of the sentences to be informed by frequency data of various sorts extracted from a corpus. The study by Millar (2011) which we cited above uses this approach: Millar's test sentences are all fabricated, but each is built around an observed non-idiomatic collocation extracted from a learner corpus.

A perhaps more straightforward use of frequency data drawn from corpora is exemplified by the eye-tracking experiments of McDonald and Shillcock (2003a, 2003b). They investigate whether the co-occurrence frequency of a pair of words (as established in a large corpus, in this case the BNC) can predict the ease of processing of the second word in that pair. The co-occurrence frequencies are expressed, in this case, as *transition probabilities*; that is, given that the first word in the pair is X, what is the probability that the second word is Y? In this case, the probability is equal to the number of times the sequence X-then-Y occurs in the BNC, divided by the total number of instances of word X – this is fundamentally very similar to a collocation calculation. McDonald and Shillcock (2003a) look at the processing of verb–object pairs, contrasting pairs where the object is probable, given the verb – e.g. *avoid confusion* – and pairs where it is less probable – e.g. *avoid discovery*. The frequencies of these bigrams in the BNC are 50 and 2 respectively, relative to 7,823 instances of the wordform *avoid* in total. McDonald and Shillcock's eye-tracking data showed that participants' eyes fixed on the object noun for a shorter time when they were reading a high-probability transition than when reading a low-probability transition. This suggests that the

(relatively) frequent verb–object collocations take less time to process than the (relatively) infrequent ones. A subsequent experiment (McDonald and Shillcock 2003b) did not look merely at one word pair in each (fabricated) sentence. Instead, the experimental sentences consisted of just over two thousand words of running text (drawn from newspaper articles), and the results were interpreted in the light of the transition probabilities across *every* bigram in the test data. The statistical analysis of these results showed that transition probability is a significant predictor of how long the eye focuses on a word (McDonald and Shillcock 2003b: 1747). That is, the length of time taken to process a word depends on how likely that word is, given the previous linguistic context; highly likely words are processed more easily, and so the eye focuses on them for a shorter time. McDonald and Shillcock thus argue that the language processor in the mind must have access to statistical information analogous to that which they themselves extracted from the BNC, and that:

> exposure to written (and spoken) language is sufficient for compiling the necessary contingency statistics . . . Experience with reading, then, represents a form of *distributional learning*: because words do not occur in isolation – rather they are encountered as part of syntactically coherent and semantically meaningful sequences – assimilating a new word into one's mental lexicon may also involve encoding its surrounding context.
>
> (McDonald and Shillcock 2003b: 1749)

Two points are notable here. Firstly, McDonald and Shillcock arrive at effectively the same perspective on how words are learnt as Hoey (2005) in the model of Lexical Priming. Hoey proposes that learning a word consists, in effect, of compiling some mental equivalent of a concordance of the word, containing information about all the contexts in which that word has been encountered. McDonald and Shillcock here propose exactly the same. We will return to the topic of the convergence of neo-Firthian linguistics with this kind of psycholinguistic study later in the chapter. Secondly, it is typical that an explanation of the observed language processing phenomenon should be, as it is here, linked to the issue of language acquisition. Language processing and language acquisition are thoroughly interwoven issues in psycholinguistics: knowing how language is processed requires thinking about how it is acquired, and *vice versa*.

The studies by McDonald and Shillcock and by Millar that we have discussed here are single examples of a vast quantity of psycholinguistic research which has found that the frequency of words, combinations of words, and structures has an effect on language processing, both production and comprehension. Ellis (2002) and Jurafsky (2003) provide extensive reviews of the research into such frequency effects on processing at many different linguistic levels, including phonotactics, orthography, lexis and inflectional morphology (Ellis 2002: 148–55; see also Gilquin and Gries 2009: 13–14). The typical result of experiments looking at frequency effects is that '[m]ore frequent categories are accessed more frequently and are preferred in disambiguation' (Jurafsky 2003: 48). It is not merely the

frequency of individual items, but also the frequencies of items in combination, that can be linked to these effects – that is, bigram frequency or word transition probability (Ellis 2002: 144; Jurafsky 2003: 50–3), as examined by McDonald and Shillcock. In short, a very large amount of frequency information of different kinds about a language is available to, *and constantly utilised by*, a competent speaker of the language; as Ellis puts it:

> [frequency information . . .] is not some idiosyncratic fact in the lexicon isolated from 'core' grammatical information; rather, it is relevant at all stages of lexical, syntactic, and discourse comprehension. Comprehenders tend to perceive the most probable syntactic and semantic analyses of a new utterance on the basis of frequencies of previously perceived utterance analyses. Language users tend to produce the most probable utterance for a given meaning on the basis of frequencies of utterance representations.
>
> (Ellis 2002: 144–5)

This does not mean that speakers are necessarily consciously aware of this frequency information, although Ellis (2002: 145) cites experiments showing that speakers are actually rather good at estimating linguistic frequencies, contrary to the usual claim in the corpus linguistic literature. Rather, it means that our subconscious, linguistic rather than metalinguistic, knowledge takes the form of probabilistic, associational information. Indeed, Jurafsky (2003: 40) argues that human thought in general 'relies on probabilistic processing'. To quote Ellis once more:

> The mechanism underlying such unconscious counting is to be found in the plasticity of synaptic connections rather than abacuses or registers, but it constitutes counting nevertheless. (Ellis 2002: 146)[1]

The frequency data that drives the experimental data reviewed by Ellis and Jurafsky can, of course, only be derived from corpora – although as Jurafsky (2003: 43–4) notes, for a long time such frequencies were typically derived only from the Brown Corpus, with the BNC and other large, modern corpora being exploited for such purposes only relatively recently, as in McDonald and Shillcock (2003a, 2003b). Gilquin and Gries (2009: 12) note that in recent experimental work, a wider range of datasets including non-prototypical 'corpora' have been exploited. Whatever corpus is used to inform studies of frequency effects on language processing, it is being treated essentially as a proxy for some hypothetical individual's entire language experience. This is, of course, a substantial simplification – the word *the* occurs some 6 million times in the BNC, but no English speaker's experience of the frequency patterns around *the* is directly equivalent to these 6 million examples – but it is methodologically justifiable in much the same way that, in corpus linguistics, the BNC may justifiably be considered a reasonable approximation to a representative sample of the English language. It is notable, however, that as soon as we consider frequency effects as quantitative abstractions from an entire lifetime's language input, we are necessarily talking

about language acquisition as well as language processing. Indeed, as we have already indicated, it is probably not possible from a psycholinguistic perspective to separate consideration of acquisition from processing: the two are profoundly interrelated. So let us now turn to a consideration of first (and, by implication, second) language acquisition.

8.2.2 Language acquisition and child language corpora

The study of child language acquisition as a field is vast, and has been undertaken from a multiplicity of theoretical perspectives. The sister field of second language acquisition, which as well as its theoretical relevance to the study of language in the mind has important practical implications for foreign-language pedagogy, is also extremely broad. For some time, theoretical work on language acquisition was dominated by the Chomskyan approach; this posits a black-box Universal Grammar as an innate module of the mind which produces a generative grammar of the adult language, given some input data in that language to work with. However, despite Chomsky's mostly negative attitude towards corpus data and methods, and the empirical study of language usage in general, child language researchers never abandoned the practice of collecting and analysing actual language data. A large quantity of foundational work in the study of child language, such as Brown (1973), was firmly based on transcriptions of children's language production. The reason for this is clear: young children lack metalinguistic awareness for a long while after they have the ability to speak, so even if we would prefer to ask them about their intuitions as a means of finding out about their language competence, we cannot. Thus, as we noted in the previous chapter, even researchers working from a Chomskyan perspective may call on corpus data when studying child language.

However, increasingly prominent as a competitor theory to Chomskyan nativism is an approach to language acquisition which is much more compatible with corpus methodologies. This approach is referred to as 'emergentist' by Ellis, who characterises it as follows:

> Emergentists believe that simple learning mechanisms, operating in and across the human systems for perception, motor action and cognition as they are exposed to language data as part of a communicatively-rich human social environment by an organism eager to exploit the functionality of language, suffice to drive the emergence of complex language representations. However, just about every content word in that previous sentence is a research discipline in itself. (Ellis 1998: 657)

Elsewhere, Ellis classifies the same approach as 'constructivist' (Ellis 2003), and as 'usage based' (Ellis 2002, 2003) – note, of course, that *usage-based* in this context is the same term, with precisely the same import, as *usage-based* in the context of functional-cognitive linguistic theory. Indeed, many psycholinguists working from a usage-based perspective adopt not only the general principles of

functional-cognitive linguistics, but also the specific descriptions and explanatory apparatus of Construction Grammar.

The explanation in usage-based approaches to language acquisition for the acquisition of constructions is, necessarily, rooted in frequency: as noted above, the psycholinguistic evidence for frequency effects in language has as much import for our understanding of acquisition as for our understanding of processing. It is important to note that the frequency of linguistic items in use does not dictate the course of language acquisition in children in a naïve, direct way. To give one basic example, the most frequent word in the spoken BNC is (unsurprisingly) *the*. If we assume the process of language acquisition to be solely determined by frequency factors, we might expect every English-speaking child to produce *the* as their first word. But this doesn't happen, and in fact very many words are learned and used in child speech before *the*. Similarly, the typical orders of acquisition established for certain features of English do not necessarily reflect the frequencies of those features in adult speech. For instance, Brown (1973) presents evidence for a particular order of acquisition of fourteen common grammatical elements in young children; but this order is not found to correlate with the frequency of these items in parental speech.

So how does frequency exert an influence in a usage-based, Construction Grammar-oriented view of language acquisition? Basically, in this approach, to acquire full language competence it is necessary to acquire a *structured inventory of constructions*, in the terminology of Tomasello (2003: 6), who gives the most comprehensive account of how this takes place. Tomasello argues (2003: 108–22) that the first pieces of language a child acquires, whether single words or multi-word utterances, are learned from the language they hear around them by exposure and repetition – that is, by straightforward associative learning. However, the multi-word utterances that children learn from their input often fall into consistent patterns of usage (for example, frames such as *Look at (something)*, *Here's (something)*, or *Let's (do something)*; examples from Cameron-Faulkner *et al*. 2003). When such patterns occur, children are able to abstract away from the individual concrete utterances to an awareness of the general schema. In other words, given a set of utterances such as *look at the ball* and *look at Teddy* in their memory, children additionally become aware of (and are able to use) the frame *look at X* – and likewise for other frames. Such a frame is, by definition, a construction: as Tomasello (2003: 100) points out, 'constructions are nothing more or less than patterns of usage, which may therefore become relatively abstract if these patterns include many different kinds of specific linguistic symbols'. But in the earliest stages, all constructions take the form of specific words with 'slots' adjacent to them – there are no fully abstract constructions such as the English transitive or ditransitive constructions. Nor are there systematic links among different constructions in the child's mind. Tomasello thus refers to these early constructions as 'item-based constructions' or 'constructional islands' (Tomasello 2003: 114–19). The structured network of adult-like abstract constructions emerges slowly by a process of further abstraction across

many different item-based constructions. For example, the transitive construction emerges from abstraction across numerous separate 'verb island' constructions with one slot before and one slot after a central verb. So the transitive construction does not exist within the child's language system until long after they have begun producing apparently transitive utterances based on these concrete verb island constructions. The process of abstraction and generalisation is driven, Tomasello (2003: 161–73) argues, by the domain-general cognitive pattern-recognition skills of analogy and distributional analysis. While Tomasello's explanation of child language acquisition in terms of Construction Grammar is perhaps the most detailed, similar views are held by many other psycholinguists; for example, in Ellis' summary:

> the knowledge underlying fluent use of language is not grammar in the sense of abstract rules or structure but a huge collection of memories of previously experienced utterances. These exemplars are linked, with like kinds being related in such a way that they resonate as abstract linguistic categories, schema, and prototypes. (Ellis 2002: 166)

The role of frequency in this model of language acquisition is that the frequency of both particular sequences of words, and particular patterns across those sequences, are major factors in the establishment of constructions, both concrete and abstract. Tomasello explains:

> In general, in usage-based models the token frequency of an expression in the language learner's experience tends to entrench that expression in terms of the concrete words and morphemes involved . . . However, the type frequency of a class of expressions (that is, the number of different forms in which the language learner experiences the expression or some element of the expression) determines the abstractness or schematicity of the resulting construction . . . (Tomasello 2003: 106–7)

Type and token frequency of any linguistic entity – including the kinds of multiword sequences we have been considering here – is ideally established on the basis of corpus data. And indeed, corpus research is prominent among the empirical evidence that supports Construction Grammar-based views of language acquisition, alongside psycholinguistic experimentation with child subjects (for example, elicitation tasks given to children to test whether they have mastered a given abstract construction – this cannot be determined solely from natural performance, as adult-like utterances can be produced using item-based constructions without an awareness of the adult-like abstraction). In particular, corpora of child language and language directed to children are invaluable in supporting specific claims within such theories of language acquisition. It is, of course, possible in some cases to use general language corpora as a first-degree approximation to the input that children receive – Pullum and Scholz (2002) make effective use of a corpus of the *Wall Street Journal* in this capacity. However, corpora of actual

child language usage data, and actual child-directed speech, open up many more options for analysis.

The creation of a corpus of child language is a major undertaking. Such a corpus is, almost by definition, a spoken corpus (the children of greatest interest to language acquisition researchers are preliterate), and as we have noted in section 1.2, recording and transcribing spoken data for a corpus is time-consuming – and thus expensive – relative to the collection of an equivalent quantity of written text. For child speech, the difficulty of transcription is greatly magnified, since many early utterances can be rather indistinct to adults who are not accustomed to the child in question's manner of speaking, and so the transcription time needed is commensurately greater. The limitations this situation places on corpus-based analysis of child language have been addressed forcefully by researchers in the field in the form of the CHILDES project (*Child Language Data Exchange System*: see MacWhinney 2000).

The CHILDES project is an initiative which collects together transcription datasets from a multitude of different studies of child language. Any researcher who has made transcriptions of child language in the course of their research can contribute their data to the CHILDES database, at which point it becomes freely available to the entire research community via the project's website. Moreover, the contributed datasets are made available in a standardised encoding format (referred to as CHAT), and an accompanying tool, the CLAN program, that is capable of processing this format and that specifically supports many of the analyses most commonly applied to child language data, is also made openly available. At time of writing, the British and American English sections of CHILDES contain the datasets created by fifty-two different projects looking at child language in use. This does not count any of the data in a variety of other languages that CHILDES has brought together and made available.

CHILDES is quite a rarity among corpora. We are not aware of any other subfield of linguistics where effectively *all* existing corpus data is available in such a unified, user-friendly and restriction-free manner. The overwhelming majority of the most significant contributions to the literature on child language has been based on some dataset which is now part of CHILDES. The implication is that the potential for replication of the findings in these studies is very high. Anyone can download any CHILDES data and check for themselves any claims made on the basis of it. Arguably, then, CHILDES sets an example which corpus linguistics (considered as a discipline) does not yet live up to (see sections 1.6.1, 3.3.3). There are two possible reasons for this. On the one hand, the difficulties of collecting child language data mean that there is a much greater penalty, in terms of effort and expense for all researchers in the field, if data reuse is not made easy. On the other hand, the extensive ethical and legal permissions that must be obtained when a researcher collects samples of child language in the first place are very likely to cover open distribution of the data – whereas corpus linguists building written corpora in particular may be given licences to copyrighted data that are much more restrictive. Be that as it may, we consider the CHILDES model

of data openness to be an extremely positive one. It is worth noting, however, that, as with field linguistics (see section 7.4), the data encoding standards and computational tools associated with CHILDES are rather different from those used in the main tradition of corpus linguistics. Many of the basic capabilities are the same: CHAT allows annotation at different linguistic levels just as corpus annotation techniques do (see section 2.3); the CLAN system possesses functions for the generation of frequency lists and keyword-in-context word or annotation searches, as does most state-of-the-art concordance software (see section 2.5). But there are some substantial incompatibilities. The user interface of the CLAN software is very different from those typical of most concordancers – even in the light of the great variety among concordancers. And the CHAT format, exemplified below, is idiosyncratic with respect to the standards familiar to corpus linguists (see section 2.3.1 for a brief discussion of textual markup):

```
*CHI:   do it again Daddy.
%mor:   aux|do pro|it adv|again n:prop|Daddy.
%xgra:  1|0|ROOT 2|1|OBJ 3|1|JCT 4|1|VOC 5|1|PUNCT
%xpho:  du it ?gein dædi
```

(An utterance from the CHILDES 'Cruttenden' database, file jane13: see Cruttenden 1978.)

The CHILDES project's impact on the field of language acquisition research is hard to overstate. Even a partial review of research using the CHILDES database would require far more space than is available here. Importantly, the data itself is of course theory-neutral, and the CHILDES database has even been used in studies that work from a Chomskyan theoretical perspective (some of which were discussed in section 7.2). But of course, the role of corpus data is much more central in the usage-based approach to language acquisition – since, as we explained above, evidence relating to frequency is very relevant to the claims made in theories such as Tomasello's. Tomasello himself (in collaboration with colleagues) has undertaken numerous investigations involving the CHILDES data. This has sometimes involved simply the study of the usage of various constructions in child language. For instance, Diessel and Tomasello (2005) use a corpus consisting of data drawn from two children, originally collected by Brown's (1973) and Bloom's (1973) studies, in an analysis of verb-plus-particle constructions in English (that is, verbs accompanied by adverb particles, such as *go away* or *turn (something) on*). Their analysis uses multifactorial statistics previously applied by Gries (2003) in an investigation of the same phenomenon in corpora of adult language. They conclude that while in some cases there is evidence for verb-particle constructions where the position of the particle is fixed, in many cases the position of the particle is determined by a combination of factors, as Gries found to be true of adult language. In other cases, the focus of investigation was the language that children hear and that is used as input for the learning of constructions. For instance, Cameron-Faulkner *et al*. (2003) look at item-based constructions in the speech of mothers to children

using data from CHILDES, showing that large proportions of such speech can indeed be characterised as involving item-based constructions – suggesting a possible means of acquisition of these structures in accordance with Tomasello's (2003) theory. Finally, data available within CHILDES has been used as a basis for devising materials for use in psycholinguistic experimentation in support of usage-based ideas. For instance, Matthews *et al*. (2004) used frequency data from the CHILDES 'Manchester' dataset (Theakston *et al*. 2001) to identify low-, mid- and high-frequency verbs. These were used in an experiment testing whether children were more likely to imitate a non-standard (SOV) word order presented with an unfamiliar, low-frequency verb than the same word order presented with a familiar, high-frequency verb. This did indeed prove to be the case – again, offering support for the notion that syntactic structures are initially item-based, and are learned piecemeal along with individual verbs.

8.2.3 Corpus data and the computational modelling of language acquisition

A further important use of corpus data in the study of language acquisition has been in another area of psycholinguistic experimentation, namely connectionist studies. Connectionist research is focused on the creation of computational models of language acquisition. What is a *model* in this sense? It is a computer program which is capable of 'learning' language – or rather, some specific aspect of language – on the basis of input data, using some quantitative machine-learning procedure, often a neural network. A neural network is a computer program which simulates a somewhat simplified view of how the brain works. A brain consists of a huge collection of *neurons* (nerve cells). Each of these neurons is connected (by *synapses*) to a large number of others, in what can be seen as a network. The basic property of a neuron is that it can 'fire', that is, send out a nerve-impulse to all the other neurons it is connected to; these neurons then either fire or do not fire in response, depending on how strong the connection to the original firing neuron is. As this effect propagates through the network of connections, hugely complex patterns of firing neurons can sweep across different areas of the brain; this is the physiological basis of thought, and the creation and continual adjustment of the networks of links between neurons is the physiological basis of learning. However, the complexity of the brain as a whole is so great that studying any aspect of learning in a controlled way is difficult – to put it informally, there is so much *else* going on in a brain that it can be difficult to see exactly how the learning is happening. A neural network can be used as a simplified model of how the brain works at this fairly low level. The computer program that runs the neural network simulates a set of nodes, each of which is connected to the other nodes in the network – but different connections have different quantitative weightings. The nodes represent neurons and the weighted connections in the model represent the links of differing strengths between neurons in the brain. Like neurons, nodes in the program's network can

be activated, and they can pass that activation on to other nodes they are connected to, depending on the weightings of the links. A set of input data is fed into the neural network, activating some nodes directly; these nodes activate others in a complex pattern across the network, and ultimately the network gives an output. In the process of training the model, the computer adjusts the quantitative weightings of the connections in the network so as to 'learn' the appropriate way to respond to different input data. As Ellis (1998: 646) puts it, '[p]sychologically meaningful objects can then be represented as patterns of this activity across the set of artificial neurons'. Crucially, a trained neural network is not limited to responding to the inputs used in training it. It can respond to new inputs using the network of connections acquired in training – this network, then, is an abstraction of the experience of the input data.

Neural networks are not actually brains and the quantitative weightings of the connections within them are not actually knowledge, in the sense that a human being 'knows' things. Yet connectionist research has shown that these simple models, which do not contain any in-advance information about the linguistic phenomenon they are being trained to handle, can often replicate quite well certain observed features of how children, with their real brains, learn language. An early example of research into language learning using such a computational model is Rumelhart and McClelland (1987), who trained a model on input data that represented the present- and past-tense forms of English verbs (regular and irregular), and showed not only that the model could learn to produce past-tense forms from present-tense forms – not on the basis of some explicitly stated rules, but simply on the basis of probabilistic knowledge created in the model by exposure to inputs – but also that the behaviour of the model, in terms of the mistakes it made, was similar to that of young children learning English. The input data (e.g. the representation of the words to be learnt) used to train the neural networks in connectionist studies is in some cases artificially devised, in isolation from actual language usage. Li *et al.* (2004: 1345) argue that this is actually a limitation of many such studies, and present an alternative approach in which the input data to the computational model comes from a spoken corpus. In Li *et al.*'s study of vocabulary acquisition, the corpus in question is a collection of child-directed adult speech extracted from CHILDES (Li *et al.* 2004: 1352). It is not simply a matter of dumping the raw running text of such a corpus into the neural network program, however. In this particular case, Li *et al.*'s input data consists of phonological representations of the words to be learnt on the one hand, and on the other hand semantic representations extracted from the corpus data by automatic processing (again computed by means of a neural network), where word meanings are derived from the transition probabilities of word sequences. So corpus data has an important role in the development of input data for connectionist research. Using textual data from a corpus, rather than isolated (possibly artificial) examples, means that the input to the models reflects more realistically the quantitative features of a child's linguistic experience. In

short, with corpus data, the computer simulation becomes a better model of the real situation.

A less radical transformation of corpus data to generate the input to a computational model (not, in this case, a connectionist neural network, but still a quantitative model) is exemplified by Monaghan and Christiansen (2010). The aspect of acquisition modelled by Monaghan and Christiansen is the ability to segment utterances into words; they utilise an algorithm that exploits previously seen examples of single-word utterances to identify word boundaries within longer utterances. Like Li *et al*., Monaghan and Christiansen use child-directed speech extracted from CHILDES, in this case extracting data from the transcripts of six separate children to compile a corpus of over 100,000 adult utterances. But unlike Li *et al*., the only transformation that Monaghan and Christiansen apply to this data, before passing it to the model, is to convert the orthographic transcription from CHILDES to a standardised phonemic representation. The resulting model proves capable, as it learns, of accurately analysing a high percentage of utterances in the corpus of child-directed speech – and although some commonly occurring units such as *isn't it*, *you tell me* and *that's right* are not segmented by the model, this actually supports theories of acquisition like Tomasello's, where such concrete constructions would be expected to be acquired as unanalysed units in the early stages. Such a compatibility between the results of quantitative computational modelling of acquisition and Construction Grammar-based theories such as Tomasello's is hardly a unique result. For instance, St Clair *et al*. (2010) use corpus data (once more, child-directed speech from CHILDES) as input into another computational model to illustrate that such a model can learn word categories (i.e. parts of speech) based on the distribution of words across highly frequent bigram and trigram contexts. This result is broadly compatible with Tomasello's (2003: 169–74) view that distributional analysis across constructions is responsible for children's learning of part-of-speech categories. And Ellis (1998) reviews a range of earlier research to argue for the mutual utility of connectionist modelling and the 'emergentist' (which for Ellis is equivalent to 'usage-based') approach to acquisition.

8.2.4 Formulaic language

One topic of increasing importance to the psycholinguistic study of language processing and acquisition is that of *formulaic language* or *formulaic sequences*. This is an alternative conceptualisation and operationalisation, common in psycholinguistic research, of the notion that we have already discussed at length in Chapter 6 under the name *collocation*. Although the specific term *collocation* is usually not preferred, this concept and that of *formulaic language* may be explicitly equated; for example, as Ellis puts it:

> Just as we learn the common sequences of sublexical components of our language, the tens of thousands of phoneme and letter sequences large and small, so also we learn the common sequences of words. Formulas are lexical chunks that result from binding frequent collocations. Large stretches of language are adequately described by finite state grammars as collocational streams where patterns flow into each other. (Ellis 2002: 155)

The parallels to neo-Firthian theory should be clear. And indeed, Ellis (2002: 155–6) and Wray (2002: 7, 13–15) both trace the general idea of linguistic competence consisting of knowledge of (semi-)precomposed sequences back to Sinclair (1991) and the Idiom Principle, as well as to other groundbreaking studies such as Pawley and Syder (1983) and, in Wray's case, to earlier theorists including Firth, Bloomfield and Saussure. On the other hand, Wray (2002: 49–50) argues for a definition of *formulaic sequence* that is rather less 'fluid' than that of collocation. In particular, the highly variable co-occurrence patterns identified by collocation procedures based solely on proximity within a window of several words around a node – which are very common in corpus linguistic analyses, see section 6.2 – do not, for Wray, qualify as formulaic.

The interest of collocation or formulaic language for psycholinguists is in its implications for an understanding of processing and acquisition. On acquisition, Ellis (2002: 156) notes that '[t]he sheer number of words and their patterns variously explains why language learning takes so long, why it requires exposure to authentic sources, and why there is so much current interest in corpus linguistics in [second language acquisition research]'; we have already discussed the role that combinatorial or sequential frequency information (e.g. word transition probabilities) has been found to play in language processing (see section 8.2.1 above; see also reviews by Ellis 2002: 157–60; Ellis and Simpson-Vlach 2009: 63). For Wray, '[t]he advantage of the holistic [formula-based] system is that it reduces processing effort. It is more efficient and effective to retrieve a prefabricated string than create a novel one' (Wray 2002: 18). Like other frequency effects on language processing, experimentation on the psychological reality of formulaic sequences is very productively informed by or devised around corpus data. For example, Ellis and Simpson-Vlach's (2009) study of formulaic language used timed reading, reading aloud and comprehension experiments (similar in kind, but different in detail, to the self-paced reading methodologies we discussed in detail above) of sentences including particular formulaic sequences. In this case, the formulaic sequences investigated were around a hundred n-grams, extracted from a range of spoken and written corpora including the Brown Family, the BNC and MICASE (see section 4.7). Ellis and Simpson-Vlach argue from their results that the psychological reality of a given n-gram – as evidenced in the degree to which it is processed quickly, as a unit – is a function, in particular, of that n-gram's mutual information score, and only secondarily of its raw frequency. Combinations such as *and at the* or *that to the*, though frequent, have a low mutual information because their component words occur frequently outside

the combination; these sequences do not act as a coherent unit in the way that n-grams such as *on the one hand* or *come into play* do (Ellis and Simpson-Vlach 2009: 74). This finding is an excellent illustration of Wray's (2002: 26–31) earlier observation that while corpus data is useful in the identification of formulaic sequences, raw frequency measurements (e.g. of n-grams) are inadequate for this purpose.

But Wray also raises other, more central concerns with the use of corpus data to identify formulae. She points out (2002: 30) that 'many word strings are indisputably formulaic, but not frequent (e.g., *The king is dead, long live the king*)'; such sequences may, in fact, not occur at all in even quite a large corpus.[2] Likewise identifying the beginning and end of a formulaic sequence may be problematic: how do we know that the word before an automatically extracted n-gram is not, psychologically, part of the sequence that this n-gram represents? More seriously, Wray identifies a difficulty in the conceptual underpinnings of the use of corpus frequencies to identify formulaicity: namely, that the frequency of a given formulaic expression ought, in many cases, to be considered relative to the frequency of all other options for expressing the same communicative function – information which is difficult or impossible to extract from the corpus. However, while these reservations about the use of corpus frequency data would appear well founded, empirically it has been the case that a very great deal of productive research into the processing of formulaic language (such as that cited above – Ellis and Simpson-Vlach 2009, or Millar 2011) has actually been successfully undertaken on the basis of such data.

Among the topics addressed by psycholinguistics, the study of formulaic language has especially significant implications for second language acquisition and, in turn, second and foreign-language pedagogy. In the light of what is known about formulaicity and other frequency effects in language, we must acknowledge that to learn a language, it is not enough to learn a lengthy list of words and a set of 'classroom grammar' rules for inflecting them and linking them together syntactically. As Ellis points out:

> For language learners to be accurate and fluent in their generalizations they need to have processed sufficient exemplars that their accidental and finite experience is truly representative of the total population of language of the speech community in terms of its overall content, the relative frequencies of that content, and the mappings of form to functional interpretation . . . the necessary representative experience for fluency must be vast indeed.
>
> (Ellis 2002: 167)

It seems clear, then, that the pedagogical problem of non-nativelike usage in learners' production cannot be fully remedied without attention to formulaic language, collocation and frequency effects in general. Earlier, we discussed the role of ideas related to Construction Grammar in certain perspectives on language acquisition. While there are many differences between the definition of constructions on the one hand, and formulae as described by Wray and others

on the other, it should be clear that from the psycholinguistic perspective it is to a large degree the same phenomenon that is being addressed. Ellis, among other authors, is happy to consider formulaic language approaches alongside Construction Grammar approaches and to treat them effectively as equivalent, or at least mutually reinforcing – see, for example, Ellis and Cadierno (2009); this convergence may be considered an instance of the broader convergence of collocation-oriented and functional-cognitive approaches to language which we will discuss in section 8.3.

8.2.5 Corpora in psycholinguistics: an afterword

So far, this chapter has looked at the role of corpus data relative to (experimental) psycholinguistics. We have seen that corpora may be exploited as a source of data concerning not only the frequency of individual elements, but also combinatorial or transitional probabilities. This data can have a crucial role in the development of methods within several different perspectives on psycholinguistics – a usage-based approach such as that of Tomasello or Ellis, a connectionist perspective, or a theory of formulaicity along the lines of Wray's. Corpora such as the BNC or Brown have been used for these purposes, but in addition a specifically psycholinguistically oriented corpus infrastructure (the CHILDES dataset, encoding standard, and tools) has emerged to support research on the central concern of language acquisition. The role of corpus data in psycholinguistics, then, seems assured. However, we would like to close by noting two points where links between corpus linguistics and psycholinguistics do not exist, or are weak, where we might expect interaction to be strong.

In the first place, given that many psycholinguists are as concerned with second language acquisition (and associated pedagogical issues) as with child language acquisition, it is perhaps surprising that there has been, relatively speaking, little interaction between psycholinguistics and the extensive research tradition based on constructing and using learner corpora (see section 4.5). The CHILDES/Talkbank data collections, for instance, contain second language learner corpora; but we are not aware of any work that combines or compares the analysis of such data with the analysis of a corpus such as ICLE. This looks like a potentially valuable avenue for researchers to explore.

Secondly, it is worth noting that nearly all the research we have cited in this chapter has been undertaken by scholars whose disciplinary background is in psychology rather than in corpus studies (an exception is Millar 2011). So, unlike the bidirectional interaction of functional-cognitive theory and corpus linguistics which we discussed in the previous chapter, the interaction of psycholinguistics and corpus linguistics is in large degree a one-way street. This point has been made in a more quantitative way by Gilquin and Gries (2009); in a literature search for papers combining corpus linguistics and experimental methods, they identify 78 per cent of the relevant research as 'psycholinguistically oriented', and only 10 per cent as corpus-linguistically oriented. Moreover, whereas psycholinguists

use corpus data in a wide variety of experimental contexts, the experimental methods used in the corpus linguistic research in their sample are more limited, with most relying on acceptability/grammaticality judgment tasks (Gilquin and Gries 2009: 13–15). One possible explanation is that some psycholinguistic methods – especially eye tracking, for instance – are avoided by corpus linguists because they require specialist technical competence which is not typically part of the training of a corpus linguist. It would be unsurprising if such (sub-)disciplinary boundaries between corpus researchers and psycholinguists had no such undesirable effects. But this cannot be the whole story, because other experimental methods (such as elicitation tasks of various sorts) do not require any such technical skills, lying within the competence of any trained linguist, but can be usefully juxtaposed with, or informed by, corpus data. For instance, Gries *et al.* (2005) describe the use of collostruction statistics to predict the results of an elicitation experiment looking at the construction in question, namely the *(verb) (object) as (object complement)* construction.

We would concur with Gilquin and Gries' judgement that it is not ideal that the interaction between corpus linguistics and psycholinguists should be driven solely by psycholinguistics. In fact, experimental methods may have much to offer corpus-based studies:

> Because the advantages and disadvantages of corpora and experiments are largely complementary, using the two methodologies in conjunction with each other often makes it possible to (i) solve problems that would be encountered if one employed one type of data only and (ii) approach phenomena from a multiplicity of perspectives . . . (Gilquin and Gries 2009: 9)

As an example of an issue in corpus linguistics which would benefit from examination from another perspective, Gilquin and Gries (2009: 17) suggest collocation statistics; while a very wide range of such statistics have been developed (see sections 2.6.2, 6.2), many of these have not been considered in the light of any kind of psycholinguistic evidence (mutual information being an exception).

Some of the literature we have cited in this section illustrates the advantage of a multiplicity of perspectives quite acutely, especially when the 'multiplicity' is more than two. This is what we would describe as *methodological triangulation*. For instance, the research by Ellis and Simpson-Vlach (2009) on the status of n-grams as psychological units triangulates by incorporating corpus data within a series of experimental investigations, as we have already outlined; but this work *also* triangulates the results against a third source of data, namely responses by instructors of English for Academic Purposes expressing their opinions on the formulaicity, cohesiveness and educational importance of the n-grams under study. The three-way methodological multiplicity allows Ellis and Simpson-Vlach to conclude (2009: 73) that 'formulaic sequences, statistically defined and extracted from large corpora of usage, have clear educational and psycholinguistic validity'. In a similar way, Gries *et al.* (2010) extend their (2005) study triangulating corpus statistics and elicitation data to a third method, self-paced reading.

8.3 The convergence of neo-Firthian corpus linguistics and functionalist linguistics

We have shown in Chapter 7, and the preceding part of this chapter, that there is increasing use of corpora (and/or frequency data derived from corpora) in a range of areas of functionalist linguistics and psycholinguistics. From the point of view of the 'corpus-as-method' school of corpus linguistics, this is both a welcome and an unsurprising development: after all, as corpus methods become more and more embedded in the day-to-day practice of linguistics, and less the preserve of computer-savvy specialists, it stands to reason that theoretical linguists of various stripes will begin using these methods. However, as explored in Chapter 6, neo-Firthian corpus linguists generally consider corpus linguistics to be a field of study of its own ('corpus-as-theory'), rather than a method to be applied in whatever other field. And as we have noted, some neo-Firthians – notably Teubert (2005) – have argued rather strongly against corpus linguistics being applied as a method within theoretical frameworks such as cognitive linguistics. In this section, we wish to argue that despite this stance, neo-Firthian corpus linguistics is showing strong signs of convergence with functionalist linguistics and psycholinguistics – and this is not merely a matter of the methods of corpus analysis being taken on board by other schools of linguistics.

In a trivial sense of 'convergence', many contemporary researchers are doing studies that are clearly both neo-Firthian and functional-cognitive in orientation. For example, Deignan's (2005) methodology is recognisably neo-Firthian (inasmuch as her analyses of collocations, units of meaning and disambiguation of lexical items in context are similar to the analytic procedures of Sinclair and others) although her work is within the theoretical framework of Conceptual Metaphor Theory. But there is a more profound sense in which this convergence is taking place; namely, that both the theoretical frameworks and the discoveries made about the nature of language are increasingly similar – and in some cases, allowing for differences of terminology, effectively identical – between neo-Firthian linguistics and one or more aspects of functionalist linguistics or psycholinguistics.

In particular, this convergence is evident when we look at the two most extensively developed neo-Firthian theories of lexicogrammar – namely Pattern Grammar and Lexical Priming. But even Sinclair's early thoughts on collocation and the Idiom Principle displayed this convergence to a degree, at least from an outside perspective. For instance, the psycholinguist Ellis, whom we cited extensively above and who views grammar from a Construction Grammar perspective, also adopts the neo-Firthian view of the lexicon as a repository of precomposed multi-word selections. Indeed, Ellis (2002) in a single paper explicitly espouses both the Idiom Principle and Construction Grammar, indicating their fundamental compatibility as seen from the standpoint of psycholinguistics.

Perhaps the most remarkable single instance of this convergence is the evident similarity between collostructional analysis (see section 7.5.1), rooted in the framework of Construction Grammar, and the neo-Firthian view on lexis and grammar whose most sophisticated expression is Pattern Grammar. But it is not merely a single point of convergence that leads us to argue for an overall rapprochement. There are in fact multiple points, major and minor, in which this kind of rapprochement can be observed, which we will now proceed to outline.

Let us consider, in the first place, the views of neo-Firthian theory and the collostruction approach on the co-selection of words and grammatical structures. From a neo-Firthian perspective, it is observed that words (and groups of words) possess tendencies to occur in particular syntactic contexts. For example, Hoey (2005: 38, 46) observes that the phrase *in winter* tends to occur in present-tense clauses, and that the noun *consequence* tends to occur as part of a complement or an adjunct (adverbial preposition phrase), but *not* as part of an object, in comparison to similar nouns. We can say, from this perspective, that *consequence* has a positive colligation with the complement and adjunct contexts, and a negative colligation with the object context. We can, alternatively, use the terminology of Pattern Grammar and look at the structure we have referred to earlier as the ditransitive, namely – from a Construction Grammar perspective – *(someone) (verb) (someone) (something)* with a meaning having to do with transfer. In Pattern Grammar, this is (one variant of) the pattern **V n n**, since clause subjects are not normally accounted as part of patterns. Hunston and Francis (1999: 88–9) list five groups of verbs, differentiated by meaning, which occur with this pattern: those 'concerned with giving someone something, or refusing to do so'; those 'concerned with doing something for someone', those 'concerned with talking, writing, or otherwise communicating something to someone'; those 'concerned with giving someone a benefit or a disadvantage'; and those 'concerned with feelings and attitudes'. These collections of verbs all have, then, a relationship of collocation or colligation with the **V n n** pattern. However, collocational associations are two-way. So, in the example drawn from Hoey, if the complement position in a clause is a colligate of *consequence*, then it follows that *consequence* is a collocate of the complement position in the clause. Likewise, in the example drawn from Hunston and Francis, if there is a co-occurrence link from *give* to the **V n n** pattern, there is a co-occurrence link from the **V n n** pattern to *give*.

This is, in virtually all details, the same kind of relationship that Stefanowitsch and Gries identify in their collostructional analysis: the attraction between slots in the syntax and particular lemmata – for example, between the verb slot of the ditransitive construction and the verb *give*. The main difference that remains between them is operational, namely that the relationship is approached beginning with the word (e.g. *give*) in the neo-Firthian approaches, and beginning with the construction (e.g. the ditransitive) in the collostructional approach. In sum, to put it baldly, the neo-Firthian study of colligation and patterns and the collostructional approach are approaching the same phenomenon from opposite ends of the link.

This judgement is demonstrated in the very similar results that the two approaches lead to. For example, all but two of the thirty verbs identified by Stefanowitsch and Gries (2003: 229) as strong collexemes of the verb slot in the ditransitive are also on Hunston and Francis' (longer) list of verbs that have the pattern **V n n**.

This convergence of concepts and of results is not surprising in view of the intellectual background of the collostructional approach. Though they ultimately choose to orient their research within the framework of Construction Grammar, Stefanowitsch and Gries (2003) explicitly situate their analysis relative to Pattern Grammar as well, emphasising many similarities between Pattern Grammar and Construction Grammar theories, and expressly equating neo-Firthian preferred terminology such as *pattern* or *idiom* to the cognitive linguistic term *construction* as synonymous notions:

> The meaningful grammatical structures that are seen to make up (most or all of) the grammar of a language are variously referred to by terms such as *constructions*, *signs*, *patterns*, *lexical/idiom chunks*, and a variety of other labels. (Stefanowitsch and Gries 2003: 210)

It is worth noting that at the same time, Stefanowitsch and Gries actually grossly overestimate the differences between the collostructional model and the neo-Firthian approach. They underestimate the scope of colligation as a phenomenon, considering it as concerned only with word-category transitions (for example, the preference of the word *the* to be followed by a noun):

> If syntax was studied systematically at all [in corpus linguistics], it was studied in terms of colligations, i.e. linear co-occurrence preferences and restrictions holding between specific lexical items and the word-class of the items that precede or follow them. (Stefanowitsch and Gries 2003: 210)

In fact, as we have noted above, colligation – at least as conceived by Hoey – is a much broader phenomenon which can include the study of those co-occurrence relationships that Stefanowitsch and Gries look at using collostructional analysis, such as the colligations of *consequence*, cited as an example above. On the neo-Firthian side, the difference is also much exaggerated: some neo-Firthian analysts have rejected Gries and colleagues' work (here and in other applications of corpus data in cognitive linguistics) in highly vituperous terms, even going so far as to deny that such research counts as the original or real corpus linguistics at all (see, e.g., Teubert 2010: 356–7). This response is explicable in light of the neo-Firthian rejection of prior theory, although as we will see in practice this rejection is not operative in some of the most sophisticated contemporary neo-Firthian accounts. Other aspects of Gries and colleagues' work which may be objectionable from a neo-Firthian perspective include its reliance on annotation, and its willingness to *begin* the analysis with an abstract construction; as we have already noted, neo-Firthian analysis typically treats collocation as bound in the first instance to individual words.

But it is clear that advances in neo-Firthian theory are introducing abstraction away from collocation as being a property strictly of word forms. Sinclair's own concept of the (extended) 'unit of meaning' is an example of such a movement away from the word as the fundamental unit of analysis. Pattern Grammar, though word-oriented in methodology, abstracts away from the word in its results – in particular, when saying which verbs have which verbal patterns, Hunston and Francis (1999) abstract from verb wordforms to verb lemmata and ignore inflections and auxiliary verb elements. And in Lexical Priming, Hoey expends much effort to demonstrate that collocations can nest and that these possibly-abstract 'nestings' can in turn be 'primed' and thus have their own collocates, colligates, semantic associations and so on (Hoey 2005: 8–11, 154–5, *passim*).

Hoey's argument for the role of nesting is highly persuasive; it is difficult to see how all the phenomena he discusses could be adequately accounted for without nesting as an explanatory factor. Yet some of the nestings Hoey proposes, particularly those that operate in such a way as to create grammatical categories such as 'noun' and 'verb', are highly generalised and thus, inherently, are abstractions across many concrete words or utterances (much, indeed, as constructions are in Tomasello's 2003, and other, accounts of language acquisition; we will return to the issue of acquisition below). If it is conceded that an abstraction can form collocations with words (or other elements; Hoey 2005: 163 includes the possibility of collocation-like 'priming' links with non-linguistic cognition such as 'feelings'), then essentially, the contrast between neo-Firthian and collostructional approaches to lexical-grammatical co-selection is voided – because that is exactly what a collostruction is: a collocation between an abstraction (the construction) and a word (the collexeme).

So when looking at the relationship of lexis and grammar, the collostructional approach of Stefanowitsch and Gries converges very greatly with the neo-Firthian approaches of Pattern Grammar and Lexical Priming. What remains are differences of methodological and terminological preference, and of thoroughness of coverage – Pattern Grammar particularly is heavily exhaustive in listing the patterns and the lexical items associated with them (Francis *et al*. 1996; Hunston *et al*. 1998), whereas work on collostructions does not as yet extend beyond a relatively small selection of lemmata and constructions. Of course, collostructional analysis is still firmly within the broader tradition of corpus linguistics, although rooted in Construction Grammar; so the question which naturally follows is whether we can see neo-Firthian theory as converging with Construction Grammar as a whole, *not just* with Stefanowitsch and Gries' application of it. Are patterns in Pattern Grammar (or, in Hoey's terms, colligations at the most abstract level) basically the same thing as constructions in Construction Grammar?

On one level the answer is plainly yes: as we indicated above constructions and patterns have long been seen as analogous. But how close is the match between them? In terms of theoretical presumptions, it is very close. Both Pattern Grammar (and other forms of neo-Firthian theory) and Construction Grammar (and other forms of functional-cognitive linguistics) see the lexicon as central to

language. Both see language production as the chaining together of items from the lexicon (constructions, collocations, extended units of meaning) where the 'joints' in the structure are 'slots' where one item can be linked into another. Are there then, perhaps, other differences? We should perhaps consider in this context what are claimed to be definitive features of the neo-Firthian side of this comparison. Hunston and Francis (1999) claim a number of differences from 'traditional' linguistic analysis established by or within Pattern Grammar:

> We have noted that corpus linguistics in general, and pattern grammar in particular, tends to cast doubt upon previous orthodoxies, such as:
>
> a. the distinction between lexis and grammar;
> b. word class as a robust system of categorisation;
> c. functional grammatical categories such as Object, Adjunct etc;
> d. constituency grammar, especially units such as group and clause.
>
> (Hunston and Francis 1999: 271–2)

As we have seen, however, Construction Grammar itself also erases the distinction between lexis and grammar, and reinterprets traditional constituency structures as constructions (which are symbolic and thus lexical). Similarly word class – that is, part-of-speech categories – has been seriously problematised in Construction Grammar, especially Croft's (2001) Radical Construction Grammar. Part-of-speech categories are seen by Croft, not as fundamental givens of the language system as in Chomskyan formalism (which we suspect to be Hunston and Francis' implicit target of contrast), but as emergent from the distribution of lexical items across particular slots in constructions. That leaves the redundancy of functional grammatical categories as a distinguishing feature of Pattern Grammar. Hunston and Francis argue (1999: 151, *passim*) that analysing patterns in terms such as object, complement or adjunct – typical terminology in functional grammar – is 'futile' on the grounds that it does not add anything to the analysis based on the surface-level features used in the description of patterns. But this may well be a function of the level of analysis at which Hunston and Francis are working. When we see grammar, and in particular verb complementation, in terms of constructions, then a notion such as 'object' is interpretable most easily as an abstraction across several things that are actually different from one another. The 'object' as a generalised syntactic role actually labels slots in a number of independent constructions, including (in English) the second nominal slot in the transitive construction and the *to*-dative, and the third nominal slot in the ditransitive.[3] These slots may be equated with one another, for example, on the basis of semantic role, since they are all prototypically patients or themes; in some languages 'object' slots can also be equated on the basis of shared morphosyntax (accusative case). But referring to these slots as 'objects' does not add anything to the analysis of those particular constructions. It rather serves to draw attention to the similarity of slots across different constructions. This may explain why Hunston and Francis do not find concepts like 'object' to be particularly valuable: making such abstract parallels between 'slots' across

patterns is not at all central to their concerns. Hunston and Francis do in fact find such functional labels highly useful as a matter of descriptive convenience, describing particular verbal patterns in such terms in many places throughout their account of Pattern Grammar. For example, their account of the **V n n** pattern (Hunston and Francis 1999: 88–90) relies on such terms to distinguish three main uses of that pattern (which, from a Construction Grammar perspective, would be considered different constructions): 'Verbs with two Objects', 'Verbs with Object and Object Complement' and 'Verbs with Object and Adjunct'.

Another point asserted by Hunston and Francis which would differentiate patterns from constructions is that patterns are *not*, in fact, abstract strings of slots into which lexical items are inserted, as their descriptive notation implies. The point of contrast here is that constructions *are* conceptualised as being abstract strings of slots. Hunston and Francis argue:

> we have talked . . . as though the pattern were a framework into which words with particular meanings could be slotted . . . This approach, however, runs counter to the work of Sinclair, for example, whose investigations . . . stress the uniqueness of each 'meaning unit'. We would come closer to the spirit of Sinclair's work if we defined a pattern as a sequence of elements including the core. For example, *approve of something* would be one pattern, *disapprove of something* would be another, *complain of something* another, *boast of something* another, and so on [as opposed to a single pattern **V *of* n**] (Hunston and Francis 1999: 86)

But in fact, Hunston and Francis' own data indicates that the more abstract 'framework-with-slots' view of patterns must be adopted in at least some cases. For example, with regard to **V n n** in the sense of 'do something for someone', Hunston and Francis note that it is impossible to list exhaustively the verbs that are used in this pattern: 'any verb that indicates an activity that you can do on behalf of someone else, or to benefit someone else, may be used in this pattern' (Hunston and Francis 1999: 90). So clearly **V n n** is not simply a descriptive convenience for multiple real patterns such as *give (someone) (something)*, *buy (someone) (something)*, *fetch (someone) (something)* and so on. Rather **V n n** really does have an actual independent existence as an abstraction across all these lexically anchored patterns, as a framework-with-slots into which almost any verb can be inserted. It is, in other words, in all details the same entity as the ditransitive construction.

Of course, Hunston and Francis' insight that such abstract patterns – whether real entities or simply descriptive conveniences – derive from collections of patterns around actual words is also in accord with usage-based theory, in particular Tomasello's (2003) view that abstract constructions emerge in children via abstraction across collections of item-based constructions. In both cases, the abstract constructions/patterns emerge from, but are not limited in use to, a set of concrete exemplar utterances.

We are not trying to argue that Construction Grammar and Pattern Grammar are identical. Rather, we are arguing that they are profoundly *convergent*. The biggest difference is the attention Pattern Grammar pays to the co-selection of lexis and grammar, which is of course a core neo-Firthian concern. This is, however, the element that Stefanowitsch and Gries add to Construction Grammar with their collostructional analyses. Once we take collostruction into account, the findings of Construction Grammar and of Pattern Grammar on the nature of language and the relationship of lexis and grammar *are* in fact very much the same. The differences of methodology and emphasis of course remain; Pattern Grammar catalogues the complexities of the syntax-semantics interrelationships with a level of detail and comprehensiveness, especially at the lexical level, which most functional-cognitive work using Construction Grammar does not reach. This may be due to the formative association of the neo-Firthian school with lexicography; in any case, it is a point on which, we would argue, cognitive approaches could productively learn from Pattern Grammar and other neo-Firthian approaches.

If the findings of the neo-Firthian Pattern Grammar and the functional-cognitive Construction Grammar are so highly convergent, do other differences remain? Many of the distinguishing features sometimes asserted for neo-Firthian research, such as the avoidance of descriptive or theoretical constructs not derived directly from the corpus (see section 6.6.4), cannot be upheld in sophisticated theories like Pattern Grammar, or indeed Lexical Priming. The notions of object, complement and adjunct in Hoey's (2005: 45) analysis do not proceed from his observations of the corpus; and we have already observed that the same notions are also prominent in Hunston and Francis' analysis in spite of the authors' own doubts regarding their utility. More substantially, Hunston and Francis are forced to import other non-corpus-driven grammatical categories such as *passive voice* or *relative clause* from pre-corpus description, simply in order to exclude these phenomena from consideration in the analysis of patterns: a relative clause is not treated as part of the pattern of the noun that it modifies (Hunston and Francis 1999: 49), and a verb in the passive voice is treated as instantiating the same pattern as the corresponding active form, even if the passive form has a different order of elements (Hunston and Francis 1999: 59–61). So it is not clear to us that, in practical terms, Pattern Grammar and Lexical Priming are materially less dependent on non-corpus-driven ideas than any of the functional-cognitive work using corpora that we have cited in the previous chapter.

Of course, from our own perspective from outside the neo-Firthian school, it is a great *strength* rather than a weakness of Lexical Priming and Pattern Grammar that, where it is productive to do so, they are willing to draw ideas from beyond the restrictions of a totally 'corpus-driven' approach. In particular, using established, uncontroversial, descriptive categories and not requiring that these be derived from scratch in interaction with corpus data is a major aid to these analyses. But, arguably, a dogmatic neo-Firthian would need to take the opposite view.

We have discussed in detail the convergence of Pattern Grammar with Construction Grammar and allied aspects of functional-cognitive linguistics. It is not by any stretch of the imagination the only such case of convergence or coincidence. We have already mentioned Lexical Priming in reference to the fact that Hoey's views on colligation are also, in large degree, convergent with the collostructional approach. But there is a more remarkable point of convergence in Lexical Priming, and that is with the psycholinguistic work which we discussed earlier in this chapter, and in particular with the work of Tomasello and Ellis, which calls on Construction Grammar just as Stefanowitsch and Gries do. The notion of 'priming', of course, is a psycholinguistic one in the first place, and one of the unique (and most valuable) features that Lexical Priming adds to neo-Firthian theory is the idea of collocation as a specifically psychological phenomenon. Going further, much of the psycholinguistic and connectionist work on probabilistic effects in language processing is highly compatible with the overall picture that Hoey (2005) gives of language competence as consisting of a web or network of item-to-item primings which are followed in production and perception. Lexical Priming is in this light a very psychologically plausible model of language. So, too, is its view of acquisition, which Hoey summarises as follows:

> language acquisition is a matter of stretches of sound stream becoming primed in such a way that they become imbued, by means of nesting, with a rich and complex web of socially embedded, genre-sensitive collocations . . . the language user becomes aware of shared primings between related words . . . [o]ut of these they will begin to abstract. (Hoey 2005: 160)

Hoey is careful to point out (2005: 161–2) that this includes the development of fully abstract grammatical phenomena such as verb complementation patterns, or even something like the passive voice. We see here an account of the acquisition of grammar very compatible with Tomasello's (2003) theory, which describes the same process in terms of constructions. Tomasello's characterisation of the emergence of abstract grammatical constructions begins with concrete instances of utterances; analysis of analogies across these instances results in item-based constructions; the same process applied to item-based constructions results in the fully generalised constructions of adult grammar. Aside from the terminology, and the separate intellectual backgrounds from which Hoey and Tomasello arrive at their respective views on language acquisition, it is difficult to see any difference here. Likewise, Tomasello sees knowledge of part-of-speech categories as emerging from distributional analysis of the occurrence of words across constructions – a view which is also found in Croft's (2001) version of Construction Grammar, as we have noted above. Hoey's (2005: 154) view that a word's part-of-speech category is the outcome of 'the combination of (some of) the word's most characteristic and genre-independent primings' chimes strongly with this

distribution-based approach to grammatical categories and their emergence in a speaker's linguistic knowledge.

Hoey characterises language competence (the end point of acquisition) as a mental concordance:

> the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context. This mental concordance is accessible and can be processed in much the same way that a computer concordance is, so that all kinds of patterns, including collocational patterns, are available for use. It simultaneously serves as a part, at least, of our knowledge base. (Hoey 2005: 11)

Again, this idea is convergent with findings from psycholinguistics that large amounts of frequency information are available to speakers as part of their linguistic competence: for example, McDonald and Shillcock's (2003b: 1749) claim, quoted in section 8.2.1 above, that 'assimilating a new word into one's mental lexicon may also involve encoding its surrounding context'. This convergence gives us some hint as to the actual shape that this mental concordance may take: following connectionist ideas, we can consider Hoey's 'mental concordance' as having the form of a distributed network of weighted links, where the weights embody a quantitative summary of all the evidence the network has ever perceived – a summary of the concordance, effectively. Gries makes a similar point:

> When corpus linguists argue against a *strict separation of syntax and lexis*, cognitive linguists agree, and many psycholinguists have long assumed that words and syntactic patterns are represented as qualitatively similar nodes in a network where, in production, lexical and syntactic nodes are activated when they fit the semantic/pragmatic meaning to be communicated.
>
> (Gries 2010b: 335)

The 'rich glossing' of the mental concordance for a non-linguistic context would then be a result of the fact that the same network that learns language also learns everything else a person knows (so that Hoey's words 'at least' in the quotation above may well be too cautious!): it is of course another convergent idea that there is no distinct cognitive system for language, only domain-general learning applied to linguistic experience. It is exciting to consider the link between Hoey's ideas and connectionist research in the light of Ellis' (2002: 146) comment that the 'counting' which contributes to the acquisition of linguistic competence is based on a mechanism 'to be found in the plasticity of synaptic connections'. We begin to see here nothing less than the tantalising possibility of tracing a link from collocation as a phenomenon in discourse to actual neurophysiological structure in the brain. We will return to this idea in the final chapter.

More cases of convergence between neo-Firthian theory and functional-cognitive-psycholinguistic theory appear regularly in the literature. Gries and

Divjak's research into behavioural profiles (see section 7.5.2) is arguably convergent with Sinclair's views on extended units of meaning and the disambiguation of polysemous words by their lexical and grammatical context, for instance. Even in the work of Teubert, a scholar notable for the strength of his adherence to neo-Firthian principles, we see evidence of the convergence for which we have argued. Namely, Teubert (2007b) points out that certain phenomena – exemplified by the complementation patterns of the noun *hatred* – cannot be described fully by Pattern Grammar without the addition of ideas drawn from valency grammar – a form of dependency grammar, which is non-neo-Firthian, and arguably within the bounds of functionalism in the wider sense. We find Teubert's (2007b) argument highly persuasive, but we have difficulty reconciling it with Teubert's earlier (2005: 4) claim that '[i]t is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question'.

It seems highly likely that work embodying this convergence will continue as functional-cognitive linguists and, perhaps especially, psycholinguists, continue to incorporate neo-Firthian concepts into their work; see, for instance, Ellis *et al*.'s (2009) and Ellis and Frey's (2009) investigations of the psycholinguistic reality of semantic prosody. This is, in our view, to be welcomed. Ideally, the convergence would be acknowledged on the neo-Firthian side, rather than being largely implicit as has been the case to date. There would be two benefits to this. Firstly, it would allow neo-Firthian scholars to draw more explicitly on valuable results and analyses in functionalist linguistics, rather than having to regenerate them from scratch. One example we would give of this, among Hunston and Francis' (1999) findings on Pattern Grammar, is that verb patterns do not include the verb's subject. In our view this effectively constitutes a rediscovery of the subject–predicate distinction, which has of course been researched in detail in non-corpus theory. For example, functionalists have investigated its relationship to information structure (topic versus comment; given versus new), and typologists have investigated its universality across a range of languages.

Secondly, we think that there are many ways in which the functional-cognitive approach to language could learn from neo-Firthian analyses. Earlier we mentioned Stefanowitsch and Gries' (2003) consideration of colligation in terms much more limited than those of Hoey. This is not the only example we could cite. There are instances of research in functional or cognitive grammar which could productively be informed by the concepts of collocation and the extended lexical unit. For example, Siewierska *et al*. (2010) undertake a corpus-based analysis of Mandarin 'splittable compounds' – a type of compound verb where other words (in particular aspect markers) can occur between the two elements of the compound – looking at the contextual conditions under which the 'split' versions are used. Siewierska *et al*. employ the approach to the study of functional variables in corpus data which we discussed in section 7.3, but the authors also spend some time discussing whether or not these splittable compounds are words, a

discussion which involves wrestling with the concept of a 'word' in Chinese. It is ultimately concluded that some of the expressions under investigation are compound words and some are phrases, and that '[splittable compounds] in Chinese straddle the morphology/syntax divide and also the structure vs. discourse one'. From the neo-Firthian perspective of the extended unit of meaning, the problem of determining whether or not Mandarin splittable compound verbs are words or not is quite simply not a problem. They are straightforward instances of extended lexical items, patterns with associated lexical items, or collocations. Siewierska *et al*. do call on the notion of collocation but go right the way back to Firth's view of the matter, without reference to more recent work – although research by Hoey and by Hunston, in particular, could clearly be invoked productively in this context. That such neo-Firthian ideas are not more widely available to functionalist analyses is a result of the distance that has existed between neo-Firthian corpus linguistics and other subdisciplines of linguistics. As functionalism and neo-Firthian theory continue to converge, it is to be hoped that this effect of distancing will decline, and ultimately disappear.

Let us summarise the argument we have made. We claim that corpus linguistics, not just of the methodologist school but of the neo-Firthian school as well, is, together with functionalist theoretical linguistics and psycholinguistics, converging on a single set of findings and theoretical postulates about the nature of language in general. This rapprochement is to be welcomed, because the very fact of the convergence occurring from very different starting points suggests that it is, to some approximation, 'the truth' about language. We would argue that the ideas on which these different kinds of linguistics are converging are, roughly, as follows:

- Language is not a unique cognitive system, it is an application of domain-general cognitive processes; the study of language is an empirical study (i.e. rejection of the Chomskyan formalist view).
- The lexicon (i.e. the mental inventory of meaningful linguistic signs: words, patterns, constructions, collocations) is the primary locus of language competence. The grammar system is either secondary to, emergent from, or even subsumed within the lexicon. At the very least grammar and lexis are ineluctably bound.
- Language processing is done by combining together meaningful elements from the lexicon.
- But the way this processing takes place is influenced and shaped, partly or in whole, by the sum total of the language user's prior linguistic experience (and associated non-linguistic experience), such that there is a strong tendency for the user to reproduce in their output, or to anticipate in their input, language similar to that which they have experienced frequently. This leads to various frequency effects including collocation in the broadest sense. This linguistic experience is also part of the language user's lexical knowledge.

- The process of language acquisition is imitative but not only imitative: specific chunks of language are learnt verbatim, but there is also abstraction from or generalisation across specific chunks of language – just as in adult language.
- Given that language competence and performance are both shaped so thoroughly by the total past experience of language in usage, the study of language use on a large scale (i.e. in a corpus) can function as an excellent proxy for the direct study of language competence – although there is no direct equivalence between any corpus and the linguistic experience of any particular speaker.
- All other things being equal, a naturally produced example of usage is better evidence than a fabricated example, and large-scale (possibly quantitative) data sampled in accordance with a well-defined corpus methodology is better evidence than isolated examples collected ad hoc. To put it another way, facts about language usage cannot always appropriately be accessed via corpus methods, but when they can be, they should be.

If we adopt these points, or some similar formulation, then the core concern becomes how corpus methods are to be combined with other methods. We are not arguing that psycholinguists or functional theorists should abandon their existing methodologies – simply that they adopt corpus methods, or the findings of corpus linguists, where possible, as the means of looking at language *in use*. Conversely, the methods and findings of psycholinguists and functional theorists should be drawn upon by corpus linguists when they go beyond looking descriptively at language in use. The need for regular and comprehensive methodological triangulation, beyond that exemplified by much of the work we have cited in this chapter, therefore becomes clear. We will consider the future prospects for such methodological triangulation in the following, concluding chapter.

8.4 Summary

In parallel to the extension of corpus-based methodologies into different areas of functionalist linguistics discussed in the previous chapter, there has been a convergence between the main findings of the functionalist enterprise and those of the neo-Firthian school of corpus linguistics. These areas of convergence include the inseparability of grammar from lexis; grammar as a phenomenon that emerges from patterns in actual usage, especially in the process of language acquisition; and the explicability of language in terms of domain-general cognitive processes (such as metaphor or category structure in cognitive linguistics, or priming in Hoey's version of the neo-Firthian idiom principle). This convergence also incorporates, and in turn informs, experimental

psycholinguistics, where investigations of language processing have increasingly shown how vital are relationships of co-occurrence among words, coinciding with and reconfirming the neo-Firthian focus on the importance of collocation.

These developments are not uniformly popular, due to the tenet held by (some) neo-Firthian scholars that corpus analysis should be undertaken independently of non-corpus-derived theories about language. However, we believe that this convergence is welcome and should be encouraged to continue. As noted, it is unlikely that a variety of approaches to language, differing in theory and in methodology, should close in on a set of points of agreement unless those points of agreement are actually, to some approximation, 'the truth' about the nature of language. Indeed, it may well be that a unified empirical linguistics – rigorously bringing together corpus linguistics, functionalist theoretical linguistics and experimental psycholinguistics – is our best chance for a comprehensive, evidence-based model of the nature of language. In the final chapter of this book, we will discuss this and other prospects for the future of corpus linguistics within the field of linguistics as a whole.

Further reading

Our argument in the latter part of this chapter has a similar thrust to that made by Gries (2010b), although Gries' perspective on the issues is naturally slightly different; Gries' paper, and much of the work to which it refers, are recommended. For other readings relevant to the second part of this chapter, we refer the reader to Chapter 6. We will concentrate here on some recommended sources either introducing psycholinguistics or looking at how corpus data can be used in psycholinguistics.

One slightly dated but wide-ranging introduction to psycholinguistics as a field is given by Garman (1990). There are many other introductory textbooks on psycholinguistics, but readers should be wary, as some are written from a strongly Chomskyan perspective (exploring language acquisition in terms of a black-box Universal Grammar) and neglect the non-Chomskyan perspective we have discussed here. For an overview of the experimental research on eye movements during reading, we suggest Rayner's (1998) review. Moving on to the link between psycholinguistics and corpus analysis, while much research has been published, book-length treatments, especially in the secondary literature, are rarer. We strongly recommend Tomasello (2003), one of the main texts we have discussed in this chapter, as he cites much work that is based on corpus-like analyses, as well as outlining an approach to language acquisition based on Construction Grammar. Several of the review articles we have cited in this chapter are highly accessible, perhaps especially Gilquin and Gries (2009). Ellis' (2002) review of research into frequency effects on language acquisition is complex but well worth the read because it summarises much of the key evidence. A later review article, Ellis and Cadierno (2009), covers much of the crucial ground on the interface of cognitive linguistics and corpus linguistics with psycholinguistic approaches

to second language acquisition (particularly in relation to Construction Grammar). We also recommend Wray's (2002, 2008) discussions of formulaic language as crucial contributions to this area. The volume on phraseology edited by Granger and Meunier (2009) includes contributions on formulaic language and collocation from many disciplinary perspectives, including those of cognitive linguistics and neo-Firthian corpus linguistics. Finally, Pawley and Syder's (1983) paper remains a very readable introduction to the issues surrounding formulaic language.

Practical activities

(A8-1) One common way that cognitive linguistics and corpus data have been combined is to compare the results of an experiment eliciting certain data to results from a corpus analysis. Here is a 'toy methodology' you could use to familiarise yourself with this approach:

- Choose three or four relatively common verbs and search for them in a fairly large corpus.
- Use either collocation statistics or the neo-Firthian collocation-via-concordance approach (see Chapter 6) to find out what collocates tend to occur immediately *after* each verb.
- Design a short questionnaire to elicit collocations based on the verbs you are looking at from the participants – there are several different ways you could do this!
- Get a handful of friends to complete the questionnaire according to their intuitions.

Did your participants come up with the same collocates or idioms that you found in your corpus analysis? Can you suggest any reasons why / why not? What potential problems are there with this kind of methodology? How might you get around them?

(A8-2) Investigate the language acquisition data to be found on the CHILDES website.

- Use the documentation and corpus file headers to get an overall estimate of the size of the 'British English' corpus *in words* (rather than in numbers of files or utterances), this being the unit most useful for comparison to other datasets.
- Download some files in CHAT format. Load them into your normal concordancer, and attempt some searches. How well does your concordance package cope with the multiple-lines-per-utterance format? Can you get reliable results from such a search? You may need to adjust some settings or preferences, depending on the software.
- If you can get this working, try to do some searches for 'baby-talk' words such as *baba*, *mama* or *dada*. How often are these said by children, and how often by adults?

Questions for discussion

(Q8-1) In some recent research, psycholinguistic experimental methods have confirmed the cognitive relevance of the mutual information statistic for collocations (see our discussion of Ellis and Simpson-Vlach 2009 in section 8.2.4). What implications does this have for how corpus linguists should study collocation? Does it necessarily mean, for instance, that all other collocation statistics (or the collocation-via-concordance method) should be abandoned? Why / why not? What further research into this area might be useful?

(Q8-2) In our discussion in this chapter, we have noted that using corpus-derived frequencies in psycholinguistic experimental design essentially involves assuming that the corpus is a 'good enough' representation of an individual speaker's complete lifetime of linguistic experience. Is this a reasonable assumption? Are there any potential negative consequences for this kind of research if this assumption is not, in fact, correct? Finally, consider whether the kind of *representativeness* a corpus needs if it is to be used as a proxy for some speaker's whole experience is the same kind of *representativeness* that was aimed for in corpora such as Brown and the BNC – that is, the corpora most often used to generate frequency data by psycholinguists!

9 Conclusion

9.1 Introduction

The preceding chapter concluded the survey of corpus methods in different fields of linguistics with which the latter part of this book has mainly been occupied. We have looked at the intersection of corpus methodologies with areas such as discourse analysis, sociolinguistics, language change, functionalist linguistics and psycholinguistics. In this final chapter, we will reflect on what we can conclude about the status of corpus linguistics within linguistics – looking at trends evident in the history of corpus linguistics up to the present time and considering how those trends are likely to continue, or, rather, how we think they *should* continue. In particular we will consider the future of corpus analysis within a framework of methodological pluralism, and the potential for corpus methods to extend beyond the field of linguistics into other areas of the humanities, sciences and social sciences. How, for example, can the methods and findings of corpus linguistics and computational linguistics continue to usefully interact? How can corpus methods be utilised in the analysis of the textually mediated world found in humanities subjects such as history, literary criticism and religious studies? And how can new methods in linguistics – for example, new approaches to neurolinguistics – inform the findings of corpus-based analyses through the process of methodological triangulation?

9.2 The story of corpus linguistics, from past to future

By surveying the variety of approaches to, and applications of, corpus linguistics over the past forty to fifty years, we have presented what may be called the 'story' of corpus linguistics. But what is the overarching theme of this narrative? In our view, essentially two broad phases in the history of corpus linguistics may be observed. The first stage, up to about the end of the 1980s, centres around the emergence of corpus linguistics primarily within two different schools of English language studies, its struggle to establish itself in the face of Chomskyan views inherently opposed to the use of corpora, and the formation of the basic set of methods and tools. The theme of the second phase, from that point

up until the present day, has been the shift in the nature of corpus linguistics as an enterprise that we have outlined in the latter part of this book. From being *in practice* a semi-independent subfield of linguistics – whether considered one in theory by its practitioners or not – corpus linguistics has become an indispensable component of the methodological toolbox throughout linguistics. The subfield labelled *corpus linguistics* that could have been coherently argued to exist in, say, 1990, is no longer so easily distinguishable from other forms of linguistics – and as we argued in the previous chapter, this rapprochement is even observable for the subset of neo-Firthian corpus linguists who *do* attempt to distinguish corpus linguistics as a separate field of linguistics.

Will there be a distinctive third phase in the ongoing development of corpus linguistics? Or, to put it more prosaically, what kind of future progressions can we predict for corpus linguistics? We anticipate that the trend of convergence between corpus linguistics and other types of linguistics will continue. The logical end-point of this development would be the extinction of corpus linguistics as a separate enterprise[1] – that is, a situation where corpus methods are simply used (where appropriate) by *all* linguists rather than being the preserve of a marginalised subgroup, as was arguably the case up until the 1990s. We do not think that this end-point can (or should) ever be reached, because even when corpus methods are fully embedded in the day-to-day practice of functionalist linguistics, sociolinguistics, discourse analysis and so on, there will still be an important role for corpus specialists whose research is concerned *with the methodology itself* – the construction and annotation of corpora, the development of new tools and new procedures, the expansion of the conceptual bases of the methodology and other such issues. The highly technical nature of some aspects of corpus methods means that not all users can be thorough specialists in the methodology; as we have noted (in Chapter 2) we do not think it is realistic to expect every linguist who uses corpus data to become fully competent in computer programming, for instance, or in the more complex statistical analyses. So we may expect a shift in what it means to be a 'corpus linguist' – from meaning someone who uses corpus data in their research, to meaning a researcher into the methodology, especially one who develops new methods and enables other linguists to apply them. To an extent this shift has already taken place, in that the research with the greatest impact in corpus linguistics is very often valued not for what it discovers about language, but for the methods it introduces or develops. To take a recent example, the findings about particular English grammatical constructions made by Stefanowitsch and Gries (2003), which we discussed in Chapter 7, are not especially revolutionary in themselves. It is, rather, *the method that these findings exemplify* – and the associated theoretical and statistical apparatus linked to collostruction – that makes this paper a key contribution to recent research in corpus linguistics. Looking further back, Sinclair's seminal (1991) text is cited more often for the illustration it gives of one prominent approach to corpus analysis than for the specific findings Sinclair presents in that work about particular linguistic items.

However, aside from the continuation of this and other current trends, there are two other directions of development for corpus linguistics which we consider both desirable and likely to come to pass. The first is that, just as corpus linguistics has become increasingly integrated as a method with other fields of linguistics, it may (and in our opinion, should) be adopted outside linguistics by other disciplines within the humanities and social sciences in particular. Secondly, the triangulation of corpus methods with other research methodologies will be an important further step in enhancing both the rigour of corpus linguistics and its incorporation into all kinds of research, both linguistic and non-linguistic. To put it another way, the way ahead is methodological pluralism. This kind of methodological triangulation is already happening, to some extent, in the case of corpus methods and the methods of experimental psycholinguistics, as we outlined in the previous chapter. But we would argue that it needs to be taken further.

The next three sections outline some particular examples of ways in which methodological triangulation or the linking of corpus linguistics to other academic disciplines (or both) may be achieved. We will look at three specific examples of new developments which have particular promise, in our view. In the next section, we will look at how renewing links between corpus linguistics and computational linguistics may allow new methods in the latter field to assist in corpus analyses. Subsequently, we will very briefly survey some recent work that exemplifies how corpus methods may be of benefit in the humanities and social sciences. Finally, we will look at what we see as the primary challenge for corpus linguistics in the future, that of methodological pluralism, and discuss some recent research which has begun to explore this.

9.3 Revisiting old friends: computational linguistics

Conspicuously absent from this book has been an extensive review of computational linguistics. Computational linguistics is the field of computer science that looks at how computer systems can be created that work with language in some way. 'Working with language' might mean, at the most theoretical level, developing computational models of the language system (see section 8.2.3); or at a very practical level, it could mean developing software that can interact with the user simply by talking to them, rather than by making the user learn to use an interface such as the window-menu-icon systems used in modern desktop software. In either case, software may be developed to analyse language input, as in speech recognition, or syntactic or semantic parsing; or to produce a language output.[2] In cases such as machine translation – automatic conversion of text in one human language to another (see Somers 2003 or Nirenburg *et al.* 2003 for an overview) – processing of *both* language input and language output is involved. Another practical problem addressed in computational linguistics is

that of extracting information from a text or texts. As a rough generalisation, we may say that understanding the information within a text is something that humans do extremely well, but slowly; computers, by contrast, do it badly but quickly. If the quality of the information extracted from text by automatic means is improved, then we can benefit in full from the speed of the computer.[3] An example of such an information extraction technique is *named entity extraction*, where the software attempts to identify all the people, places, organisations and so on mentioned in a text. When techniques like this are applied across a large set of texts, it is often referred to as *text mining* (see Feldman and Sanger 2007), which is one example of a more general problem called *data mining* – the identification of patterns and extraction of information across very large datasets. A priority in the field of text mining is currently *biomedical* text mining (see Cohen 2010) – that is, extracting information from large collections of text (usually academic papers) on biology or medicine. The amount of scientific literature being continually produced in biology and medicine is now far beyond any individual scientist's ability to keep up with more than a fraction of it; one goal is for text mining software to automatically produce accurate summaries of this vast complex of material for biologists and doctors.

This very short and admittedly incomplete overview of computational linguistics has hopefully served to illustrate that while there are some substantial overlaps with corpus linguistics, it is fundamentally a separate field. Corpus linguistics is ultimately about *finding out about the nature and usage of language*. While computational linguistics may also be concerned with modelling the nature of language computationally, it is *in addition* focused on *solving technical problems involving language*. Computational linguistics as a field converged greatly with corpus linguistics in the 1980s and 1990s (see McEnery and Wilson 2001 for an overview; see also section 4.3) as methods based on corpus data became an essential part of many areas of computational linguistics, and, likewise, advances in corpus annotation enabled by innovations in computational linguistics (such as probabilistic tagging) proved of great value for linguists working with corpora. However, this period of intersection seems to have passed, to some degree, except perhaps in the relatively narrow areas of part-of-speech (POS) tagging, parsing and other forms of tagging such as the annotation of anaphoric relations (see Botley and McEnery 2000). Computational linguistics still makes extensive use of corpora and other sorts of digital 'language resources'. Indeed, many of the very largest corpora are constructed mainly for use in computational linguistics, for example the English Gigaword corpus,[4] as well much of the other material assembled by organisations such as the Linguistic Data Consortium and the European Language Resources Association. But apart from these shared resources, there is now relatively little interaction between the concerns of the two fields.[5]

We consider this unfortunate, in light of the benefits that earlier cross-fertilisation has had. POS tagging is a near-indispensible tool for linguists' corpus searches, for instance. Likewise, the development of parallel corpora was in part

driven by the usefulness of such data for machine translation (for example the CRATER corpus, see section 1.7), though parallel and comparable corpora have proven of equal interest for contrastive analysis of languages (see Borin 2001; McEnery and Xiao 2007a, 2007b). However, certain more recent advances in the computational analysis of language have not 'crossed over' to corpus linguistics and become commonly used methods in this way, as will be discussed shortly. Computational linguistics is an 'old friend' of corpus linguistics, in that they have been and continue to be linked (not least, perhaps, by the somewhat inaccurate perception of those outside the fields of a greater similarity between them than actually exists); but it is a friendship which needs to be renewed and reinvigorated if both sides are to get the most out of the link.

It is, of course, impossible to foresee in full detail all of the possibilities for linguistic analysis that may emerge from a reinvigoration of this connection. However, some predictions may be made based on the nature of research that *has* continued at the intersection of corpus linguistics and computational linguistics, and by considering potential applications to linguistics of computational techniques that have recently emerged. Much of this work is fundamentally semantic in nature. Research based on semantic tagging, in particular, is notable for having taken place in both disciplines. It has, for instance, been applied as an approach to searching for metaphorical language in corpora (see Koller *et al*. 2008; Semino *et al*. 2009). But it has also been applied to problems of information extraction, for example by Sawyer *et al*. (2002), who use semantic tagging to address the problem of extracting from documents *about* a piece of software an account of the requirements for that software. We may expect the importance of semantic tagging as a method in corpus linguistics to continue to grow as the implications of the analyses it permits become clearer (Rayson 2008 makes some substantial initial steps in this direction).

Another field of recent interest in computational linguistics, namely *sentiment analysis* (also know as *opinion mining*: see Liu 2010 for an overview), takes the computational analysis of meaning far beyond the level of individual words or phrases. The main aim of sentiment analysis may be characterised roughly as the automatic identification of what a writer *feels* about the topic of the text they are writing (or, alternatively, their *opinion* of that topic matter). Often, this is directed at the practical task of searching for documents on the web that express a particular opinion, such as blog posts or customer reviews of products in online stores. Since about 2001, this field of computational linguistics has expanded substantially (Pang and Lee 2008: 5–7). Automated sentiment analysis is a non-trivial task for many reasons. For example, whether a text expresses a positive or negative opinion is not necessarily deducible from the number of positively or negatively evaluative words or expressions present in the text. Linguistic phenomena such as pragmatic implicature allow a negative opinion to be conveyed without any straightforward negative expressions such as *bad*, *awful*, *very poor*, *I hate X* and so on being used (see Pang and Lee 2008: 17–22 for examples).

The key point for our purposes is that sentiment analysis has had little or no impact on the field of corpus linguistics, in spite of some fairly obvious uses for it in discourse analysis and pragmatics. The kind of evaluative language that sentiment analysis looks at *has* been addressed within (corpus) linguistics (see particularly Hunston and Thompson 2000), but to date this strand of research has remained a somewhat separate undertaking to sentiment analysis. The fault for the lack of integration is not solely on the side of the linguists; Pang and Lee (2008: 13–15) identify applications for sentiment analysis in business and government intelligence-gathering, political science, law and sociology – but make no mention of linguistics. But the potential utility of sentiment analysis for linguists of various kinds is clear. It would be very useful for researchers in pragmatics to be able to search for texts which express a positive or negative subjective opinion. Enabling such searches for a standard corpus such as the BNC will, of course, entail finding solutions to problems beyond those inherent in sentiment analysis itself. For instance, in a general corpus, what is the appropriate unit across which sentiment should be analysed? The datasets of typical interest for sentiment analysis are divided into individual texts that are clearly appropriate units (e.g. individual blog posts or individual product reviews). It is much less obvious that a single text in the BNC, which may be tens of thousands of words in extent, can or should be meaningfully analysed for its 'sentiment'. But on the other hand, neither is it clear that the sentence or the paragraph (or, in speech, the utterance) would be more suitable units. Nevertheless, the potential value of this approach as an extension to the corpus linguistic toolbox of methods is illustrative of what linguistics has to gain by an active re-engagement with cutting-edge research in computational linguistics.

9.4 The textually mediated world: the humanities and social sciences

The social and cultural world that we as human beings inhabit exists, and is expressed and recorded, to a very large degree by means of language. Socially and politically, we are overwhelmingly concerned with matters (such as right, wrong, friendship, love, justice, freedom, the law, ownership, desires, grudges) that exist only as abstractions within our minds and within the discourse that links our individual minds together. Even non-linguistic cultural experiences such as visual art or music are encountered in the context of a huge quantity of talk and writing *about* these experiences. Similarly, any knowledge of our culture's past beyond personal experience exists principally as linguistic (typically textual) communication – and the same is, of course, necessarily true of any aspect of the culture that refers to our future. So it can be said that experience of the human world is largely a textually mediated experience, and, to that extent, human beings live in a textually mediated world.

This is probably the most important reason why many disciplines within the humanities and social sciences are, to a very large degree, concerned with the study of texts. In literary studies, the object of investigation extends from literary texts to critical texts about that literature. The study of religion includes, very prominently, the study of sacred texts and both ancient and modern commentaries on them. History as a discipline is partly based on the physical evidence of objects from the time period under study – which may themselves include text, such as inscriptions on coins or monuments – but also on the study of textual records from the period, or later texts that discuss it. Similar arguments could be made for most other areas of the humanities. All these fields have developed modes of analysing text directed at the particular requirements of the field – as, indeed, have linguists. But for the most part, these analytic techniques are targeted at *individual* texts. In these subject areas much as in linguistics, however, the information technology revolution of the late twentieth century has meant that very large amounts of text of potential research interest are available in machine-readable form. For example, very large bodies of literary texts are available in a number of online databases. It seems clear that the methods developed in linguistics for handling corpora can be applied to help humanities scholars extend their hand-and-eye techniques of analysis to these much larger bodies of text. In fact, we would argue that any field that is based, primarily or in part, on the study of text can benefit from corpus methods in any research context where the body of text that is of interest expands beyond the point where hand-and-eye methods of analysis can fully encompass its contents.

To give a very basic example, a literary critic who is looking at a single novel may have no pressing need for corpus methods. But if that analyst wishes to look at all the literature of a period, they may well find that corpus methods have something to offer them – not as a replacement for, but rather as an addition to, their existing analytic techniques, just as linguists have found. In writing a study of, for instance, the Victorian novel in English, the researcher's own experience of critically reading examples of the Victorian novel will necessarily be at the heart of the study. But there is no reason, given the availability of corpus methods, that such a study should be limited to those novels the researcher has read. A corpus of hundreds of Victorian novels could be assembled from text freely available on the web, and used to supplement the core critical analysis. For example, if the researcher wishes to make the claim that the treatment of a certain topic by – say – Dickens is of particular note, then this must obviously be supported by a reading of the Dickens text in question. But it could be reinforced and extended by examining, as a point of comparison, a corpus search for instances of that topic (using a wordform search or a semantic tag search) in the works of *all other authors in that period*.

Humanities research exploiting corpus tools and resources is a subset of the field of *humanities computing*, or *digital humanities* as it is often known nowadays (see McCarty 2005). Digital humanities research includes the development and exploitation of many forms of database, not just corpora. For example, work has

been done to create databases of images (Bailey 2010) and of archaeological objects (Heath 2010). Even when work in the digital humanities is oriented towards text, it does not necessarily treat the text from a primarily corpus-oriented perspective. One central concern is the creation of electronic critical editions of particular documents (see Deegan and Sutherland 2009), where a searchable text may indeed be created, but enabling corpus-type analyses such as concordances, collocations and so on is not the main goal.

However, the sheer size of many of the textual databases that are now available means that corpus methods – and, in particular, tools developed with the aim of analysing the very largest corpora – have a great deal to offer humanities research. We might consider, for instance, the *Early English Books Online* (EEBO) database,[6] which includes as far as possible everything published in print form in English before 1700. This amounts to many thousands of documents and hundreds of millions of words of text. Indeed, EEBO is larger than any but the very biggest corpora of present-day English. EEBO is obviously an incredibly valuable resource for historians and literary scholars. However, the native search tools of the EEBO interface are oriented to the individual-document approach to text analysis. Searches of the EEBO database for a particular term produce *not* a set of instances of that term, but rather a list of documents in which that term appears (each of which can then be viewed individually as scanned graphics or as a text file). The utility of such a database could obviously be greatly extended by supplementing (not replacing) these tools with the facilities available in most good modern concordancers (see Chapter 2) such as concordances, collocations, n-grams, keywords, dispersion or distribution analysis, query thinning and so on. Further gains can be achieved by applying corpus annotation techniques to enhance the power of searches. Many corpus annotation procedures have been adapted to historical text, including POS tagging (Rayson *et al*. 2007) and semantic tagging (Archer *et al*. 2003). These may require a pre-process to add regularised, spellings as one of the layers of annotation (see Pilz *et al*. 2008; Baron *et al*. 2009),[7] in the light of the marked variability in English orthography in earlier centuries.

But if corpus techniques are to be used at all – let alone embraced – across a breadth of humanities subject areas, it is clearly important that the corpora should be made accessible via a user-friendly interface. It is absolutely unrealistic to expect the majority of humanities scholars to get to grips with the theoretical and methodological underpinnings of corpus linguistics as a field – as laid out in the early part of this book – before making use of corpus tools in their analysis. It is similarly unrealistic to expect them to learn computer programming, or the statistics associated with corpus data, or the operation of corpus software tools which are not entirely self-explanatory. As we noted in Chapter 2, these are the directions in which corpus tools are developing in any case, in order to make the methods accessible to all linguists. The necessity of this trend is only that much greater when we consider researchers in other fields.

This level of user-friendliness and accessibility has not yet been achieved. For this reason, so far, studies which are identifiable as applying specifically corpus-based methods to questions in fields other than linguistics have often been done either by, or in collaboration with, specialists in corpus linguistics. Our own research as methodologists has exemplified some work of this kind. McEnery (2005) uses corpus-based analyses to look at historical and sociological, as well as linguistic, aspects of bad language in English; Hardie and McEnery (2009) address text reuse and the expression of bias in early Modern English journalism; Gregory and Hardie (2011) explore procedures by which historical corpora can be mined for geographical information, and the resulting datasets presented in visual form as maps.

However, it is to be hoped that as corpus methods become more accessible to the non-linguist, it will become possible for research like this to be undertaken by humanities scholars without methodological support. The challenge for corpus linguists is to enable this shift to take place.

9.5 The challenge ahead: integrating corpora with new methods in linguistics

As we outlined above, we see methodological pluralism, and methodological triangulation, as critical to the future agenda of corpus linguistics. This argument is, of course, not original to us. We discussed pluralism between corpus methods and experimental psycholinguistics in the previous chapter; but that is not the only context where the case for methodological triangulation has been made. For example, Hollmann and Siewierska (2006) present a rationale for tempering corpus results with other techniques when approaching dialectology. For instance, since most available dialect corpora are small, it is difficult to study relatively infrequent phenomena in them (Hollmann and Siewierska's example is the ditransitive construction). However, speaker elicitation tasks can be employed to produce additional data which, critically, can be *combined together* with the corpus data. This is a key point of a methodologically pluralist approach: one type of data does not necessarily trump another, rather different types of data may be used to complement one another – either by confirming some finding, or conflicting with it. In such an approach, multiple types of data may be used to shed light on a single phenomenon. Another example of this kind of approach is provided by Hoffmann and Lehmann (2000), who use questionnaires to test whether or not speakers are conscious of low-frequency collocations (fewer than one hundred instances in the BNC) which nevertheless are highly statistically significant. They are thus able to show that speakers can 'memorise combinations that are only encountered approximately 5 times over a whole year' given

an assumption that the BNC is sufficiently representative of speakers' linguistic experience (Hoffmann and Lehmann 2000: 17, 31). Similarly Arnold *et al*. (2000), in their corpus-based study of constituent ordering which we reviewed in section 7.3, additionally perform an elicitation experiment which shows that, in addition to the factors identified in their corpus data, a speaker's personal preference has a role to play in determining the order of noun phrases in a clause. More such work is summarised by Gilquin and Gries (2009). Gilquin (2006) provides an example of triangulation where one method problematises the conclusions we might draw from the results of another; she demonstrates that the corpus frequency of the senses of a word does not, in at least some cases, correspond to the prototypicality of those senses as established by methods of cognitive linguistics. These studies exemplify the kind of methodological triangulation that is already being undertaken with corpus data, and, in fact, a substantial proportion of the literature this book has reviewed could be argued to be (potentially or actually) methodologically plural in this way. In this section, we would like to consider the prospects for *new* instances of methodological triangulation with *other* kinds of linguistic research, taking as our example the field of neurolinguistics.

Neurolinguistics may be defined as the study of the language system as it exists in the physiology and neurology of the brain. Core issues in neurolinguistics include: Which parts of the brain are involved in language processing? What are their functions? How do they interact? Neurolinguistics has come a long way since the widely known early work on language in the brain, where the language disorders (aphasia) suffered by people with particular kinds of brain damage – identified via autopsy – were studied in order to deduce which areas of the brain are involved in language. This technique was used in the nineteenth and early twentieth centuries to identify the linguistic functions of key areas of the cerebral cortex, most notably Broca's area, shown to be important in the production of fluent speech, and Wernicke's area, shown to have a role in the association of words and meanings. A model of language processing called the Wernicke–Geschwind model, based on this kind of investigation, was prominent for some time (Geschwind 1974). However, there are obvious drawbacks to attempting to model typical language processing in the brain based on disordered language systems and post-mortem study of damaged brains. More recently, neurolinguistic research has been based on brain imaging technology, which allows researchers not only to study language directly in the living brain, but also to use evidence from the study of non-disordered language systems in the development of new models (see Ingram 2007: 59–64). For example, a type of scan called an fMRI (functional magnetic resonance imaging) allows variations in blood flow in the brain to be monitored. This indicates which parts of the brain are 'working hard' while a particular task is undertaken, because blood is directed to whatever parts of the brain have the greatest requirements for energy (and thus oxygen) at any given moment. Using an fMRI to observe blood flow during particular tasks in language processing and production allows a profile to be constructed of how different parts of the brain are involved in different types of linguistic thought.

Of course, neurolinguistics does not stand in isolation from other areas of linguistics. For example, those views on language processing in the brain that treat it as a *modular* process (following Fodor; see Fodor 1983) have drawn inspiration ultimately from Chomskyan theory (Ingram 2007: 79). Other links have been drawn between neurology and linguistic theory. For example, Feldman and Narayanan (2004) argue on neurolinguistic grounds that the core semantics of a word is *embodied*. What is meant by this is perhaps most easily understood by a consideration of Feldman and Narayanan's main example, the verb *grasp*. It has been shown that the same groups of cells in the brain are activated both when someone is performing the action of grasping, *and* when witnessing someone else grasp something. This phenomenon is also observed in non-human primates; the nerve cells in question are referred to as 'mirror neurons'. What Feldman and Narayanan suggest is that this same neural substrate is the locus of the meaning of the word *grasp*. That is, the process of comprehending *grasp* is a neurological simulation of the action of grasping. This argument is also extended to metaphorical uses of *grasp*, with particular reference to the type of abstract-as-concrete metaphors considered important in Conceptual Metaphor Theory. Metaphorical expressions such as *to grasp an idea* or *to grasp an opportunity* are also understood by neurally simulating the physical action. Feldman and Narayanan link this theory explicitly to the Construction Grammar view of the relationship between form and meaning; they thus create a variant of this theory called Embodied Construction Grammar, where the meaning of constructions is linked specifically to the neural structures underlying the production and perception of physical entities around us (i.e. embodiment as the core aspect of meaning). This theory has important implications for how neurolinguistic evidence should be interpreted; for instance, '[o]ne should not expect language to be any more (or less) localized [in the brain] than other perception and action' (Feldman and Narayanan 2004: 391).

But it is fair to say that, so far, there have not been any significant attempts to link neurolinguistics and corpus linguistics specifically. This is, to a degree, not surprising; at first glance, there seems to be a gaping abyss between the study of text in a corpus on the one hand, and the study of brain scans on the other. But we would argue that in conceptual terms there is actually a great deal of similarity. Our fundamental interest is the *nature of language*. If we assume that language does indeed have the ontological status of a cognitive-neurological system as well as a social phenomenon (as we have noted in Chapter 8, some corpus linguists argue that it does not), the ultimate object of study is the same in both corpus linguistics and neurolinguistics, and in both cases there is a fundamental commitment to empiricism. Furthermore, in both corpus linguistics and neurolinguistics, we study the language system not directly but by observation of epiphenomena – in the case of corpus linguistics we look at the language system's output on the large scale, in the case of brain imaging we look at either the language system's blood-flow requirements or some other physiological feature associated with it. So if corpus linguists and neurolinguists have, at least in part, the same object

of study and the same underlying epistemology, methodological triangulation between these fields is, arguably, not only possible but very highly desirable.

Indeed there are certain areas where the findings of corpus linguistics quite evidently require integration with neurolinguistics. Collocation is prominent among these. What does a collocate look like in the brain? What patterns of activity may be observed when collocations are being produced or processed, and how does this differ (if at all) from the activity observable when language functioning by what Sinclair calls the Open-Choice Principle is being processed? Linguistic and psycholinguistic theories about lexicogrammar – whether corpus-oriented, like Hoey's theory of Lexical Priming, or not – all make certain claims about what is stored in the memory of a speaker. What are the reflexes of these stored entities in terms of the ultimate physiological substrate, that is, interconnections among neurons in the brain? For instance, how does learning a collocation, or being 'primed' to use Hoey's term, act to change the actual structure of the brain? An answer to this and similar questions will inevitably require an integration of corpus evidence and neurological evidence from brain imaging and other techniques. It is not yet clear how such an integration may be undertaken, although it seems clear that psycholinguistics will play an important role in building the link; but it is in our opinion one of the most exciting vistas for future research.

Exactly the same argument could be made for methodological triangulation with psycholinguistics, different forms of functional and cognitive linguistics, and so on. However, as we outlined in the previous chapter, methodological triangulation between corpus linguistics, functional-cognitive linguistics and psycholinguistics is already happening, as is triangulation of corpus linguistic and sociolinguistic methods (see Chapter 5), albeit with much scope for further work; whereas linking corpus findings to neurological findings is clearly a challenge for the future. If this challenge is met, however, the prospect opens up of the development of what we might call a Unified Empirical Linguistics, where evidence of all kinds – textual, psychological and neurological – is as a matter of course used in concert to uncover the fundamental nature of language. It is in the context of such a Unified Empirical Linguistics, we would argue, that corpus linguistics will reach its full potential as a methodology.

9.6 The final word

With this chapter, we conclude our admittedly incomplete survey of corpus linguistics, and its intersection with other types of linguistics and other fields of study. If we have accomplished nothing else, we hope to have given some indication of the very great breadth of corpus linguistics. There are few areas of linguistics where there is no possible role for corpus methods (the most obvious example being Chomskyan theory), and there are increasingly many where corpus methods have become central. At this point, then, it is safe to say

that the place of corpus analysis in linguistics is assured, and that (as we have illustrated in this chapter) the directions in which it seems likely (or desirable) that the field will develop can only lead to yet further expansion of its scope – to new questions, new academic disciplines and new methodological combinations and integrations. Above all, it is this remarkable diversity in the uses of corpus data – and the diversity of viewpoints that these uses may promote – that makes corpus linguistics as a field so utterly refreshing.

Glossary

Please note: the entries in the glossary are intended to serve as a simplified *aide-memoire* for the terms listed; in the main text, we discuss many of them in detail, in some cases identifying problematic aspects of the terms which are *not* repeated here. For a broader set of terminological definitions, see Baker *et al*. (2006).

adjunct. A noun phrase or preposition phrase which adds additional information about the state-of-affairs in a clause, without being one of the *arguments* (or *participants*) of the verb.

American Structuralism. A school of linguistics, dominant in the USA in the first half of the twentieth century, which focused on the analysis of how elements are structured at a series of linguistic levels: phonology, then morphology, then syntax and so on.

Americanisation. The process whereby non-American dialects of English become more like American English over time.

anaphora. An anaphor is a pronoun or noun which refers to an entity that it does not specifically name – often something that has already been mentioned in the preceding discourse.

anaphoric annotation. The process of tagging all instances of anaphora in a corpus to indicate precisely *which* noun phrases in the preceding discourse they are referring back to.

annotation. (a) Codes within a corpus that embody one or more linguistic analyses of the language in the corpus. (b) The process of adding such information to a corpus.

anonymisation. The process of deleting names of persons and other potentially identifying elements from a corpus text, usually a transcription of a conversation.

argument. A noun phrase or preposition phrase which indicates one of the participants in the state-of-affairs of the clause. This includes the subject, object, indirect object and so on. In traditional grammar, the argument noun phrases are often said to be *compulsory* elements of the clause syntax.

aspect. A grammatical category, often inflected on verbs or indicated by auxiliary verbs, which indicates the internal time-structure of an event.

attributive adjective. An adjective that modifies a noun within a noun phrase (compare *predicative adjective*).

auxiliary verb (or just *auxiliary*). A category of grammatical words, sometimes considered a type of verb, which do not indicate any particular 'action' but instead mark features of the tense, aspect or modality of another verb (which is often called the 'main' verb).

balance. A property of a corpus (or, more properly, of a corpus sampling frame). A corpus is said to be *balanced* if the relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled.

balanced corpus. See *sample corpus*.

belles lettres. A genre in the Brown Corpus sampling frame, consisting of texts considered 'literary' (in the sense that their main purpose is deemed aesthetic rather than practical) that are not fiction – primarily essays on various topics, often relating to art and culture. The term is now archaic except as a description for this sampling frame category.

cataphora. Like *anaphora*, except that the cross-referring noun or pronoun refers to an entity that *has yet to be mentioned* and that is introduced *subsequently* in the discourse.

central modal. In English grammar, one of the nine historically oldest modal auxiliary verbs: *can*, *could*, *may*, *might*, *shall*, *should*, *will*, *would* and *ought*. Formally, two features set them apart from other verbs used to mark modality: they do not take the *-s* suffix in the third person singular, and they are followed by an infinitive main verb without the *to* infinitive marker (except *ought*).

chi-square. A statistic used for significance testing. It is calculated by summing the squares of the differences between each value in a table and the corresponding expected value (what the observed value would be if the variables tabulated don't make any difference).

Chomskyan linguistics. A school of linguistics which arose from, and largely succeeded, American Structuralism, and which takes its cues from the work of Noam Chomsky; it can be characterised as a formalist approach that seeks to identify the most abstract logical systems underlying the rules of language (considered as an innate human capacity). Chomskyan linguistics was the most dominant school of linguistics in the 1960s and 1970s. See also *generative grammar*.

client/server software. An approach to software design where a task is split across two different programs; a client which interacts with the user and translates and transmits the user's requests, and a server which actually does the work of carrying out these requests before transmitting the results back to the client. The client and the server can run on the same machine, or on different machines, in which case they communicate across a network or the Internet. The typical example of a client/server is a web browser (client) and website (server).

cluster analysis. A statistical technique which identifies groups of objects in datasets where many different variables exist; the similarities among objects are calculated using *all* the variables.

clusters. Either (a) an alternative term for n-grams or (b) the data groupings produced by cluster analysis.

Cognitive Grammar. A major approach to grammar within cognitive linguistics, associated primarily with Ronald Langacker.

cognitive linguistics. An approach to linguistic theory, closely allied to functionalism, which seeks to explain language in terms of what is known about how the mind works (cognition).

collexeme. See *collostruction*.

colligation. A co-occurrence relationship between a word and a grammatical category or context.

collocation. A co-occurrence relationship between two words. Words are said to *collocate* with one another if one is more likely to occur in the presence of the other than elsewhere.

colloquialisation. The process whereby a language changes over time to become more 'speech-like', usually by virtue of grammatical or lexical features that are associated primarily with speech becoming more common in all kinds of text.

collostruction. A co-occurrence relationship between a grammatical construction (the *collostruct*) and a lemma that tends to occur in one of its slots (the *collexeme*).

comment. One of the elements into which the content of a sentence or clause can often be broken down; the *topic* is the part that indicates what the clause is about and the other part, the comment, says something about the topic.

comparability. Two corpora or subcorpora are said to be *comparable* if their sampling frames are similar or identical. For example, a corpus of 1 million words of English news text and a corpus of 1 million words of French news text are *comparable* for the purpose of contrasting English and French; an English news text corpus and a French spoken corpus are not.

comparable corpus. A corpus containing two or more sections sampled from different languages or varieties of the same language in such a way as to ensure comparability. If more than one language is involved, this is a type of multilingual corpus; contrast *parallel corpus*.

complement clause. A subordinate clause that acts as the syntactic complement of a noun or verb. In English, clauses beginning with *that* are often complements: *I know [that this is the case]* (complementing a verb); *The idea [that the sky is pink] is crazy* (complementing a noun).

computational linguistics. The field of research applying computer science techniques to language and language data. This includes, but is not limited to, various kinds of research using corpora, such as text mining. (Despite

the name, computational linguistics is very often in practice a branch of computer science rather than a branch of linguistics.)

Conceptual Metaphor Theory (CMT). A theory that suggests that commonly used 'dead' metaphors actually reflect the use of metaphor as a mode of thought, where the way we think about and talk about some (usually abstract) target domain is structured on the basis of some (usually concrete) source domain.

concordance. A display of every instance of a specified word or other search term in a corpus, together with a given amount of preceding and following context for each result or 'hit'.

concordancer. A computer program that can produce a concordance from a specified text or corpus. Most concordancers today can also perform other types of analyses.

connectionism. An approach to psycholinguistics which seeks to model language learning or processing by training neural networks, i.e. computer programs that loosely model the networks of neural cells that exist in the brain.

connotation. The *suggested* meanings of a word, as compared to its *denotation* (what it directly means and refers to). Connotations are emotional overtones associated with a word and can be positive or negative. For example, *tyranny* has negative connotations, *warmth* has positive connotations.

consistency of annotation. Corpus annotation is said to be consistent if decisions on ambiguous, borderline or theoretically controversial cases (e.g. in part-of-speech tagging, whether participles should be tagged as adjectives or verbs) are made in the same way every time.

constituent. In syntax, another term for a phrase such as a noun phrase, verb phrase or clause. A *constituency analysis* indicates where each phrase begins and ends and how phrases are nested within each other; this is one of the main approaches to syntactic *parsing*.

Construction Grammar. A theory of grammar within cognitive linguistics, where all syntactic structures and idioms are considered to be *constructions* – meaningful units, stored in the mental lexicon, which may consist of concrete words as well as abstract slots; in this view language is produced by combining together words and constructions, linking them via the slots in the constructions.

corpus construction. The process of designing a corpus, collecting texts, encoding the corpus, assembling and storing the relevant metadata, marking up the texts where necessary and possibly adding linguistic annotation.

corpus-based linguistics. Depending on the author, may mean either (a) any approach to language that uses corpus data and methods, or (b) an approach to linguistics that uses corpus methods but does not subscribe to the principles of the so-called *corpus-driven* approach.

corpus-driven linguistics. Depending on the author, may mean either (a) a neo-Firthian approach to corpus linguistics, or (b) a corpus method that is entirely bottom-up rather than top down.

corpus-informed linguistics. A term sometimes used for linguistic research which uses corpus data but does not aspire to *total accountability* to the data in the corpus.

co-text. The text surrounding a word or phrase of interest. The word *context* is often used with this meaning. However, some writers reserve *context* for non-linguistic context (e.g. the situation in which it was written) and thus use *co-text* to refer to linguistic context.

count, non-count. Count nouns can be individuated, and thus (in English) can be made plural, can occur with the indefinite article and so on; they usually refer to items. Non-count nouns cannot; they often refer to substances. Examples of non-count nouns in English include *sand*, *rice*, *water*. Nouns can be both count and non-count in different contexts.

Critical Discourse Analysis (CDA). The study of *discourses*, in the sense of language practices that embody, express or construct some ideology or worldview, as exemplified in texts. CDA takes an explicitly sociological and political approach to the study of discourse.

data-driven learning. A way of using corpora in language teaching that involves the learners being given direct access to the corpus and a tool for searching it, the intention being that their exploration of the corpus helps their learning of the language.

democratisation. A process of language change whereby forms that explicitly mark social relations of unequal power are avoided and fall into (relative) disuse.

diachronic. Relating to the study of a language or languages as they change over time. A diachronic corpus samples texts across a span of time or from multiple time periods.

digital humanities. A field of study that uses computer data resources (often, but not always, textual in nature) to address research issues in the humanities and arts, such as literary criticism, cultural studies or history.

discourse. In the most basic sense, a discourse is a stretch of language longer than a single sentence. By extension from this, the term has a range of other meanings: an entire text; the whole of a population of texts; or a way of using language or way of thinking about the world. See also *Critical Discourse Analysis*.

discourse prosody. See *semantic prosody*.

disfluency. An irregularity in spoken language production that typically produces a sentence that, according to traditional grammar, would not be considered well-formed. Disfluencies include false starts, utterances broken off halfway or reformulated halfway through, slips of the tongue,

fillers such as *um* and *er*, and other phenomena arising from the unplanned, spontaneous nature of spoken language.

ditransitive. The ditransitive is a grammatical structure where a verb is linked to two objects, the direct object and the indirect object, as well as a subject. This ditransitive construction, of the form *(someone) (verb) (someone) (something)*, alternates with the *to*-dative construction *(someone) (verb) (something) to (someone)*. A ditransitive verb is one that occurs in these two structures, for example *give* or *send*.

emergentist. A view of language acquisition where grammar emerges from experience of language via domain-general learning processes, rather than being an innate function of the mind or a mental module that is independent of general cognition.

encoding. The process of representing a text as a sequence of characters in computer memory. When discussing corpus encoding, we may also consider issues of *markup* and *annotation*.

ergativity. A grammatical pattern is described as *ergative* if it treats the subject of an intransitive clause in the same way as the object of a transitive clause (with the subject of a transitive clause being treated differently). This is the opposite of the *accusative* pattern found in English and most European languages.

ethics. In the context of research, ethics refers to the standards developed in academia for what is considered acceptable and unacceptable with regard especially to the treatment of participants in the research.

extended unit of meaning. In neo-Firthian theory, a linguistic element with a single meaning, part of the lexicon of the language, which is not necessarily limited to a single word: it may instead consist of the collocations, colligations and semantic preferences and prosodies around a central (node) word.

factor analysis. A statistical technique for analysing a dataset with lots of variables; it groups the variables that behave similarly together into a smaller number of factors.

falsifiability. A scientific hypothesis or claim is *falsifiable* if it is possible, at least in principle, for evidence to be found that would demonstrate that the hypothesis is not true.

Fisher's exact test. A statistical significance test often used as a better alternative to the chi-square test and the log-likelihood test.

formulaic sequences. Fixed or semi-fixed sequences of words, assumed to be stored as single units in the mental lexicon and not analysed into sub-units. This term is often used by psycholinguists; in other areas of linguistics, the terms *idiom*, *collocation*, or *multi-word unit* are more common.

frequency list. A list of all the items of a given type in a corpus (for example, all words; all part-of-speech tags; all four-word sequences) together with a count of how often each one occurs.

functionalism. An approach to linguistic theory which seeks to explain the forms of language structures by reference to how they are used – involving such factors as communicative purposes, how utterances are processed and so on.

generative grammar. Any theory of grammar which aims to define a set of formal rules that can generate all and only the grammatical sentences of a given language or of all languages. These theories of grammar are inspired ultimately by the work of Noam Chomsky.

genre. See *register*.

given information. Words in a sentence that refer to entities or events that have already been introduced into the discourse in the preceding text, and are thus assumed to be in the speaker and hearer's minds, are described as containing given information.

grammaticalisation. The process whereby, typically over a period of centuries, content words can lose their main meaning and become grammatical elements (usually accompanied by phonetic reduction). In English, for example, all auxiliary verbs that mark aspect or modality in the modern language are derived from what were once lexical verbs that could stand alone as the main verb of the clause.

head. In the syntactic analysis of phrases, one word is usually identified as the *head* or main word. The grammatical properties of this word determine the grammatical properties of the phrase (for example, an adjective phrase behaves grammatically much like a lone adjective, a noun phrase like a lone noun).

headword. See *lemma*.

HTML. *Hypertext markup language*; the system of encoding used on the World Wide Web to indicate the structure and formatting of a webpage as a hypertext document, using tags in <angle> <brackets> added to a plain text file. A form of *SGML*.

hypertext. Text containing links that can be followed to navigate around a document or between documents. The World Wide Web is essentially a massive collection of hypertext.

IBM-compatible personal computer (PC). The formal term for what is normally just called 'a PC'. Nowadays, this covers most computers other than those manufactured by Apple and the biggest servers, mainframes and supercomputers.

idiom. A phrase or other multi-word unit whose meaning cannot be deduced simply by combining the meanings of the words within it; most theories of language agree that for this reason, idioms and their meanings must be part of the lexicon of the language. See also *collocation*, *extended unit of meaning*, *formulaic sequence*.

Idiom Principle. The idea in neo-Firthian theory that most language is produced and comprehended by linear chaining-together of idiomatic elements drawn from the lexicon (these elements may be dubbed *idioms*,

collocations or *extended units of meaning*). In this view, only a minority of language is produced or comprehended by the contrasting Open-choice Principle, where individual words are built into clauses according to abstract rules of grammar.

indexing. The process where a computer program prepares an *index* of a text or corpus that it can later use to search for words or phrases in the corpus without going sequentially through the whole text, or even without accessing the original text at all.

intransitive. An intransitive verb is one that does not take a direct object. The intransitive construction is the grammatical structure, found in many languages, of a clause containing a verb and its subject but no objects.

introspection. A linguist who relies on their own *intuition* as a source of data about the nature of their language is said to be using introspection. These intuitions often take the form of judgements about whether a given sentence is grammatical in their language. Relying solely on introspection is not considered an appropriate approach to language in corpus linguistics.

intuition. See *introspection*.

key word in context. A format for displaying a concordance where the search result is lined up in a central column, and the columns on either side contain a short chunk of the context preceding and following each result in the corpus. The standard abbreviation is KWIC. 'Key word' here means the search term, not a *keyword* in the more usual sense.

keyword. A word that is more frequent in a text or corpus under study than it is in some (larger) reference corpus, where the difference in frequency is statistically significant.

L1. Customary abbreviation for *first language*, referring to the native language or mother tongue of a speaker.

L2. Customary abbreviation for *second language*, referring to a language learnt subsequent to early childhood. Any language other than the first is referred to as an L2 even if it is the third or fourth to be learnt.

lemma. A group of wordforms that are related by being inflectional forms of the same base word. The lemma is usually labelled by that base or stem. So, for instance, in English *destroy*, *destroys*, *destroying* and *destroyed* are all part of the verb lemma *destroy*; but the noun *destruction* is a separate lemma, because it is related to *destroy* by derivational rather than inflectional processes. The notion of a *headword* (as found in a dictionary) is generally equivalent to that of *lemma*.

lemmatisation. A form of corpus annotation where every token in the corpus is labelled to indicate its *lemma*.

lexical bundles. See *n-grams*.

lexical item. Either (a) a general term for a *lemma* or anything else found in the mental lexicon, or (b) in neo-Firthian theory, another term for the *extended unit of meaning*.

lexicogrammar. An approach to language which sees the words of a language and its grammar as closely linked. Some version of this position is accepted by a very wide range of schools of linguistics, the main exception being *Chomskyan linguistics*.

lexis. The words and other meaningful units (such as *idioms*) in a language; or, the study of these units. The lexis or *vocabulary* of a language is usually viewed as being stored in a kind of mental dictionary, the *lexicon*.

log-likelihood test. A significance test similar to the chi-square test, but generally considered more reliable, especially when working with small values.

machine-readable text. Text represented as sequences of characters encoded as numbers in computer memory or saved in a disk file. Image files are not machine-readable in this sense.

mainframe computer. A very large, powerful computer; most organisations would have only one or a very few. In the era before ubiquitous personal computers, corpus processing would typically be done on a research institute's mainframe.

manual annotation. The method of corpus *annotation* where the analytic codes are added to the text by a human being.

markup. Codes inserted into a corpus file to indicate features of the original text other than the actual words of the text. In a spoken text, markup might include utterance breaks, speaker identification codes and so on; in a written text, it might include paragraph breaks, indications of omitted pictures and other aspects of layout.

markup language. A system or standard for incorporating *markup* (and, sometimes, *annotation* and *metadata*) into a file of machine-readable text. The standard markup language today is *XML*.

metadata. Data *about* data; in a corpus, this usually means data about the texts – for example, the author, date of publication, title and source of a written text, or information about the sex, age and social class of speakers in a spoken text.

metalinguistic knowledge. A speaker's explicit knowledge *about* their language (as opposed to their *linguistic* knowledge, which is largely implicit, and is the actual procedural knowledge that allows them to produce and comprehend utterances in that language).

modal verb. An auxiliary verb which marks some feature of modality on another, main verb. English has two groups of verbs like this: the *central modals* and the *semi-modals*.

mood (or *modality*). A grammatical category marked on verbs; different types of modality indicate permission, obligation, possibility, ability, necessity and other notions of this type.

monitor corpus. A corpus that grows continually, with new texts being added over time so that the dataset continues to represent the most recent state of the language as well as earlier periods.

multi-dimensional (MD) analysis. An approach to studying types of text based on using factor analysis to identify dimensions of variation across a large collection of texts; associated primarily with Douglas Biber.

mutual information. A statistic that indicates how strong the link between two things is. Mutual information can be used to calculate collocations by indicating the strength of the co-occurrence relationship between a node and collocate.

neo-Firthian. A label for the tradition of corpus linguistics based on the work on John Sinclair, who applied the ideas of J. R. Firth to corpus analysis. See also *corpus-driven linguistics*.

neural network. See *connectionism*.

neurolinguistics. The study of how language works in the brain (part of the nervous system, hence *neuro-*).

new information. Words in a sentence that introduce entities or events that have not previously been mentioned before are described as containing new information.

n-grams. An n-gram is a sequences of *n* elements (usually words) that occur directly one after another in a corpus, where *n* is two or more. Studying n-grams (also called *clusters*, or *lexical bundles*) is one way to operationalise the analysis of *collocation*.

node. In the study of collocation, the *node* word is the word whose co-occurrence patterns are being studied. If we look at a list of collocates of *cheese*, for instance, we are treating *cheese* as the node word.

normal distribution. A pattern observed in many datasets where most of the values are close to the average (mean) value, forming a 'bell-shaped' curve when this is plotted on a graph. Many statistical procedures assume normal distribution, but this can be problematic since language data such as word frequencies is typically not normally distributed.

normalised frequency. A frequency expressed relative to some other value, as a proportion of the whole – for example, frequency of a word relative to the total number of words in the corpus. Normalised frequencies can be compared even if they arise from datasets of different sizes.

oblique. In the grammar of noun case, *oblique* refers to all case markers other than those that indicate core grammatical roles such as subject and object.

Open-Choice Principle. See *Idiom Principle.*

operating system. The most basic program running on a computer that controls its hardware and makes available a platform for other programs to be built on. Microsoft Windows and Unix are the two most commonly encountered operating systems.

optical character recognition (OCR). The process of automatically generating machine-readable text from a computer file containing an image (scanned or photographed) of a printed page.

orthographic transcription. When an audio recording is transcribed orthographically, each word of speech is transcribed in its standard

spelling. There is no attempt to indicate how the word is actually pronounced, as there is in a phonemic transcription.

parallel corpus. A corpus consisting of the same texts in several languages. This typically means a set of texts written in one language together with each text's translation into a second language (or into several other languages).

parsing. The process of analysing the syntactic structure of a text or part of a text (such as a sentence). By extension, any kind of corpus *annotation* which indicates syntactic structure.

participle. A non-finite verbal form, normally one that can function as an adjective. English verbs have two participles; the present participle is used in marking progressive aspect and the perfect or past participle is used in marking perfect aspect and the passive.

part-of-speech tagging (POS tagging). The process of adding part-of-speech tags to a text; a form of *annotation*. Usually undertaken automatically by a *tagger* program.

part-of-speech tags (POS tags). Codes that can be added to each word in a corpus to indicate the grammatical category of that word (e.g. noun, verb, adjective, etc.).

passival. A now-archaic English grammatical structure where a non-passive progressive verb is used with the meaning of a passive progressive: for example, using *the meal was cooking* to mean the same as the passive clause *the meal was being cooked*.

passive. A grammatical construction where the normal *arguments* of a verb are rearranged: the direct object becomes the subject, and the subject is demoted to an optional adjunct. In English, the passive is formed by auxiliary *be* followed by the past participle, and the demoted subject is marked with the preposition *by*: *This house was built by an architect*. The normal, non-passive form is called the *active*.

perfect aspect. In English, the perfect aspect is a grammatical category applied to verbs, indicating a completed event; it is formed from the auxiliary *have* followed by a past participle, for example *he has done it*.

phonemic transcription. A transcription of a spoken text where the actual sounds (phonemes or *segments*) produced by the speakers are represented using the International Phonetic Alphabet. A *phonetic transcription* is similar but goes beyond the phoneme level to indicate detailed phonetic features of each segment in context.

population. The complete set of 'things' that a sample is trying to represent. In corpus linguistics, the population is the entirety of a given language, or of a given variety of language. We attempt to make generalisations about a population, say of *all* English newspaper language or of *all* spoken British English, based on a corpus that contains only a *sample* of that population.

postmodification. Elements in a noun phrase that modify the head noun and follow it (in English, relative clauses and preposition phrases).

predicate. In the grammatical analysis of a clause, the predicate is the whole of the clause except the subject – that is, it includes the verb and any other noun phrases or other constituents.

predicative adjective. An adjective that is a verbal complement and is thus linked to an argument of the verb, but that is not part of the noun phrase of that argument. For example, in *Joey was tall*, *tall* is a complement of the copula verb, and describes the subject *Joey* while not being part of the subject noun phrase.

premodification. Elements in a noun phrase that modify the head noun and precede it (for instance, other nouns, adjectives or genitive noun phrases, as in *Mr Smith's lovable mountain dog*).

present participial clauses. In English, subordinate clauses centred around a present participle whose implied subject is the same as the main clause they are linked to; for example *Derek walked down the street [whistling a little tune]*.

priming. A psychological phenomenon where a word is understood more quickly if a semantically related word has been perceived beforehand; the speaker is said to be 'primed' for the second word by the first word. The idea is extended to collocations by Michael Hoey's theory of Lexical Priming.

progressive aspect. An aspect in English which indicates an event that is ongoing and not bounded in time. It is formed with auxiliary *be* followed by the present participle, for example *she is doing it*.

prosodic annotation. A form of corpus *annotation* applied to a spoken text, which indicates features of the prosody such as pauses, intonation and so on. It can be applied to either orthographically or phonemically transcribed text and must currently be created manually.

prosody. In phonetics, any feature of the pronunciation of an utterance other than the actual *segments*, typically covering things like intonation patterns, tone groups, pauses and so on. The term sometimes (as in Firth's usage) covers changes to segments depending on their context, i.e. what is otherwise called assimilation (e.g. nasalisation of segments near to a nasal consonant).

qualitative analysis. An analysis that does not rely on numeric data. For example, if you look at a word in a concordance and analyse its usage based on your own understanding of the examples you have seen, this is a qualitative analysis. The best corpus analyses are both qualitative and quantitative.

quantitative analysis. An analysis that is based on statistics, whether basic statistics such as frequency, or more advanced techniques such as significance testing, cluster analysis or factor analysis.

register. A way of classifying texts according to non-linguistic criteria, such as the purpose for which a text was produced, the intended audience, the level of formality, whether its purpose is narration or description and so on. For

most purposes, the term *genre* can be considered equivalent, though some authors differentiate them.

relative clause. A clause which modifies a noun within a noun phrase, like an adjective does. It contains a relative pronoun which has the same referent as the noun it modifies. In English relative clauses follow the head noun, e.g. *[Some people [who I know]] came to visit*.

relative frequency. See *normalised frequency*.

replicability. Hypotheses or claims based on the findings of scientific investigations are verified by attempts to replicate them by repeating the procedure of the original investigation; a claim that proves not to be *replicable* is discarded.

representativeness. A *representative* corpus is one sampled in such a way that it contains all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety that it samples. See also *balance*.

sample. A single text, or extract of a text, collected for the purpose of adding it to a corpus. The word *sample* may also be used in its statistical sense by corpus linguists. In this latter sense, it means a group of cases taken from a population that will, hopefully, represent that population such that findings from the sample can be generalised to the population.

sample corpus. A corpus that aims for *balance* and *representativeness* within a specified *sampling frame*. See also *snapshot corpus*, contrast *monitor corpus*.

sampling frame. A definition, or set of instructions, for the samples to be included in a corpus. A sampling frame specifies how samples are to be chosen from the population of text, what types of texts are to be chosen, the time they come from and other such features. The number and length of the samples may also be specified.

segment, segmental phonology. A segment or *phone* is a single speech sound such as a consonant or vowel; segmental phonology is the side of phonology that focuses on the behaviour of segments rather than supra-segmental features like tone or prosodic features.

semantic association. See *semantic preference*.

semantic preference. A co-occurrence pattern between a word and a semantic category of words. The co-occurrence with any given member of the category may not be especially frequent (in contrast to *collocation*). But the category counted as a whole *does* co-occur frequently with the word.

semantic prosody. The tendency exhibited by some words or idioms to occur consistently with either positive or negative meanings.

semantic tagging. A process of corpus annotation where each word is assigned one or more tags indicating its semantic category, or in some cases its semantic relationships to other words.

semi-modal verb. In English grammar, a category of auxiliary verb constructions which are used to mark modality but which do not have all

the grammatical features of the nine *central modals*. Examples are *be going to*, *have to*, *want to*.

SGML. See *XML*.

significance test. A mathematical procedure to determine whether a result is *statistically significant*.

snapshot corpus. A corpus which contains a fixed sample representing a specified form of the language at a specified time; a type of *sample corpus*. Contrast *monitor corpus*.

stance adverb. An adverb which indicates the attitude of the speaker towards a state-of-affairs (rather than the manner in which that state-of-affairs occurs, which is the basic use of adverbs). For example, in *Hopefully Derek will win the race*, the adverb *hopefully* does not indicate that Derek will win the race in a hopeful manner; it indicates that the speaker has an attitude (stance) of hope towards the possibility of Derek winning the race.

standardised type–token ratio. A variation of *type–token ratio* that averages the ratio across equal-length subsections of a corpus. This produces a standardised statistic that can be compared across corpora, even if the corpora are not of the same size. See *type–token ratio*.

stand-off annotation. Corpus annotation is typically encoded in the same text files as the actual original text, but this need not be the case. When tags are stored in separate files, and a computer program is used to link text to tags when needed, the annotation is described as *stand-off*.

statistical significance. A quantitative result is considered statistically significant if there is a low probability (typically less than 5 per cent) that the figures extracted from the data are simply the result of random chance, and do not indicate what they seem to indicate. A variety of statistical procedures can be used to test statistical significance.

structure dependency. A feature of grammatical rules, namely that they always refer to structure and never to sequential order. In Chomskyan theory it is argued that the universality of structure dependency is evidence of the innateness of language.

subjunctive. A grammatical form, very marginal in English, that can be taken by verbs, normally in contrast to an indicative form, as part of the system for indicating *modality*.

synchronic. Relating to the study of a language or languages as they exist at a particular moment in time, without reference to how they might change over time. A *synchronic corpus* contains texts drawn from a single period – typically the present or the very recent past.

synthetic negation. In English grammar, negation by means of a word other than the actual negative marker *not*: for example using *nobody* instead of *somebody/anybody* or marking a noun phrase with the article *no*.

tagger. An informal term for a computer program that automatically applies some type of analytic corpus *annotation*. On its own, the word *tagger*

usually implies a part-of-speech tagger, but lemmatisers and semantic taggers are also types of tagger.

tagging. An informal term for corpus *annotation*, especially forms of annotation that assign an analysis to every word in a corpus (such as *part-of-speech tagging*, *lemmatisation* or *semantic tagging*).

text. As a count noun: a text is any artefact containing language usage – typically a written document (book, periodical, leaflet, sign, webpage, t-shirt slogan) or a recorded and/or transcribed spoken text (speech, broadcast, conversation). As a non-count noun: collected discourse, on any scale.

text mining. In computational linguistics, the study of extracting information from large amounts of textual data (i.e. corpora) via automated analysis; for example, identifying all the places and people mentioned in a body of text.

textual markup. See *markup*.

theoretical linguistics. The branch of linguistics concerned with the structure of language as a system – phonology, morphology, syntax and semantics in particular – and with the different theories that have been proposed regarding the nature of this system.

***to*-dative**. The English grammatical construction of the form *(someone) (verb) (something) to (someone)*, e.g. *the architect gave the money to the shopkeeper*. It is distinguished by the recipient or indirect object being marked by the preposition *to* and coming *after* the direct object, and alternates with the *ditransitive* construction, of the form *(someone) (verb) (someone) (something)*.

token. Any single, particular instance of an individual word in a text or corpus. Compare *type*, *lemma*.

topic. One of the elements into which the content of a sentence or clause can often be broken down; the topic is the part that indicates what the clause is about and the other part, the *comment*, says something about the topic.

total accountability. The principle that, when we study some phenomenon in a corpus, our investigation must include *all* relevant data in the corpus, without excluding any examples that may be problematic for the analysis.

transition probability. For a sequence of two words (or part-of-speech categories, or any other linguistic units), the transition probability is the probability that, given an instance of the first of the pair, the following word is the second of the pair.

transitive. A transitive verb is one that has both a subject and a direct object. The transitive construction is the grammatical structure, found in many languages, of a clause containing a verb together with its subject and direct object.

treebank. A parsed corpus.

t-test. A statistical test sometimes used in the calculation of *collocations*.

type. (a) A single particular wordform. Any difference of form (for example, spelling) makes a word into a different type. One type may occur many

times in a text or corpus; all tokens that consist of exactly the same characters are considered to be examples of the same type. See also *token*, *lemma*. (b) The term *type* is also used in corpus linguistics with its standard meaning, for example when discussing *text types*.

type–token ratio. A measure of vocabulary diversity in a corpus, equal to the total number of types divided by the total number of tokens. The closer the ratio is to 1 (or 100 per cent), the more varied the vocabulary is. This statistic is not directly comparable between corpora that are of different sizes.

typology. A branch of linguistics which investigates the structures of the languages of the world, how these can be classified, and what the trends and patterns detected across languages can tell us about the nature of language.

Unicode. A standardised system for encoding machine-readable text that allows all the writing systems of the world to be represented (unlike earlier systems, which could typically only handle one or two writing systems at a time).

usage-based. A term sometimes applied to views of linguistic theory or psycholinguistics in which actual discourse is considered an important explanatory factor (i.e. cognitive, functionalist or emergentist views).

wordform. See *type*.

XML (eXtensible Markup Language). A *markup language* which is the contemporary standard for use in corpora as well as for a range of data-transmission purposes on the Internet. In XML, tags are indicated by <angle> <brackets>. *SGML* (Standard Generalised Markup Language) is an older, very similar standard with a slightly different set of rules for creating tags.

Notes

1 What is corpus linguistics?

1 See Chapter 2 for a more detailed discussion of concordance software.
2 For an interesting case study of this sort of problem, see Hardie *et al*. (2006: 226–9) and Hardie (2007).
3 See www.publications.parliament.uk/pa/cm/cmhansrd.htm.
4 See www.ucl.ac.uk/english-usage/projects/ice-gb/.
5 See http://sourceforge.net/projects/thedrs.
6 At the time of writing, this corpus is 400 million words in size. It is available for use, free of charge, from www.americancorpus.org/. The corpus expands by 20 million words per year.
7 This form of coding was used in the LOB corpus (Johansson *et al*. 1978).
8 We will explore the criticisms made of annotations in corpus texts in section 6.6.3.
9 There do exist some cases of full-depth CDA studies done on the scale of a large corpus, for instance Baker *et al*. (2008).
10 Our discussion here is based in part on two earlier publications, McEnery and Xiao (2007a, 2007b).
11 An introduction to the EMILLE project can be found at www.emille.lancs.ac.uk.
12 See Frankenberg-Garcia (2008) for an example of the use of the COMPARA corpus to study lexical frequency in translated and non-translated material.

2 Accessing and analysing corpus data

1 Ashby (1981), Tiersna (1982), Sun and Givón (1985), Schiffrin (1985), Mohan and Zader (1986), Biber (1986), Du Bois (1987), Fox (1987), Besnier (1988), Clark and Carpenter (1989), Gropen *et al*. (1989), Biber and Finegan (1989), Youmans (1991), Birner (1994), Hudson (1994), Rickford *et al*. (1995), Baayen and Renouf (1996), Roberts (1998) and Mougeon and Nadasdi (1998).
2 Carden (1982), Di Sciullo *et al*. (1986), Downing (1993), Siewierska (1993), Berg and Abd-El-Jawad (1996), Paradis and Lacharité (1997) and Sampson (1997).
3 Bloom (1990), Déprez and Pierce (1993), Hyams and Wexler (1993) and Snyder and Stromswold (1997).
4 For a recent example of this same process, consider the work of Breton *et al*. (2008) on gravity and neutron stars.
5 Geoffrey Leech, personal communication.

6 EAGLES proposed standards in a range of areas (see www.ilc.cnr.it/EAGLES/home.html); for the guidelines on part-of-speech and syntactic annotation, see respectively Leech and Wilson (1994, 1999) and Leech *et al*. (1995).
7 See Gries (2010c) for a disarmingly honest discussion of the limitations faced by corpus linguists who are not also computer programmers.
8 The GATE system was developed at the University of Sheffield and is in wide use. See http://gate.ac.uk/.
9 There are 12,387 instances of 了 in the LCMC; this word can be a particle or an auxiliary and is used mainly to mark the perfective aspect.
10 Our thanks to Oliver Mason for giving us details of the early search programs used at Birmingham.
11 While we are treating it as a third-generation concordancer, Xaira (and the SARA system it descends from) also have features of the fourth generation. See www.oucs.ox.ac.uk/rts/xaira/.
12 *Collocations* may be defined for the moment as links between pairs of things that tend to co-occur in a corpus; but see the extended discussion of collocation in Chapter 6.
13 *Keywords* are words that occur relatively more often in a text or corpus being analysed than they do in some reference corpus, where the difference in frequency is statistically significant.
14 See www.ucl.ac.uk/english-usage/resources/icecup/.
15 See http://pelcra.ia.uni.lodz.pl/ for PELRCA and http://hnc.ilsp.gr/en/ for the Hellenic National Corpus.
16 Corpora available within this system include the Corpus of Contemporary American English, the Corpus del Español and, as mentioned earlier, the BNC.
17 See http://cwb.sourceforge.net/.
18 See http://cwb.sourceforge.net/demos.php for a number of examples of such use of CQP.
19 Regular expressions are queries that use a well-established, fairly standard and extremely powerful search syntax that was originally developed within the field of computer science. The most notable feature of this syntax is the wide range of 'wildcards' that are available, some of which have been borrowed by simpler search languages. For example, in a regular expression, a dot stands for 'any character' and a question mark makes something optional; so the regular expression query *walk.?.?.?* would find *walk*, *walks*, *walked* and *walking* – that is, all the inflectional forms of the verb *walk* – but also the surname *Walker*.
20 An example is the Digital Replay System, which like some of the previously mentioned tools is open-source; see http://sourceforge.net/projects/thedrs/.
21 See http://ucrel.lancs.ac.uk/claws/trial.html.

3 The web, laws and ethics

1 See, for example, Véronis (2005).
2 See www.baal.org.uk/dox/goodpractice_full.pdf.
3 See www.darpa.mil/about.aspx.
4 Reader's letter to the UK newspaper *The Independent*, 19 August 1999.
5 See http://blog.veronis.fr/2005/02/web-googles-missing-pages-mystery.html.

4 English Corpus Linguistics

1 It later became the *International Computer Archive of Modern and Medieval English*.
2 The eponymous *ICAME* journal, established in 1979, and available online at http://icame.uib.no/journal.html.
3 See www.rodopi.nl/senj.asp?SerieId=LC. Note, however, that not all ICAME conferences gave rise to books in this series. Leitner (1992) and Nevalainen *et al*. (2008) are two examples of collections of papers from ICAME that were published elsewhere.
4 See http://ice-corpora.net/ice. For an example of the research results produced using the ICE corpora, see Bolton *et al*. (2003).
5 See www.lancs.ac.uk/fass/projects/corpus/LCMC/.
6 See http://catalog.elra.info/product_info.php?products_id=84.
7 See www.emille.lancs.ac.uk/.
8 See Meyer (2002: Chapter 4) for an engaging overview of the various approaches taken in ECL to the production of treebanks.
9 It should be noted that Scott moved to Liverpool from Lancaster rather than Birmingham.
10 Of course, corpora of the academic writing of L1 speakers writing in their L1 also exist – see, e.g., Ebeling and Heuboeck (2007).
11 See www.uclouvain.be/en-cecl-lindsei.html.
12 See also the excellent online bibliography of learner corpus research at www.uclouvain.be/en-cecl-lcBiblio.html.
13 Quoted from www.pearsonlongman.com/dictionaries/corpus/learners.html, last accessed December 2010.
14 Quoted from www.cambridge.org/elt/corpus/learner_corpus2.htm, last accessed December 2010.
15 See www.linguistics.ucsb.edu/research/sbcorpus.html.
16 See Adolphs (2008).
17 Douglas Biber, personal communication.
18 See section 5.2 for a brief description of this corpus.
19 See www.ldc.upenn.edu.
20 See http://corpus.byu.edu/.

5 Corpus-based studies of variation

1 For a fuller discussion of this point, see section 1.4.4.
2 For a discussion of sampling regimes, see section 1.4.3.
3 See also our discussion of Biber (2004) in section 5.3.2.
4 See http://icame.uib.no/brown/bcm-los.html for a list with the details of each sample included in the Brown Corpus.
5 Clauses expressing commands, requests, etc. which use the present subjunctive, with verbs such as *suggest* or *ask* introducing the clause, are analysed as mandative subjunctives. The following is an example from the BNC (file FBJ, sentence 911): *He asked that his son **be** informed*.
6 This corpus, made freely available by Mark Davies at http://corpus.byu.edu/time/, is a 100-million-word corpus of written American English built from articles that have appeared in *TIME* magazine. The material in the corpus covers the period from 1923 to the present.

7 Biber has also used the longer term 'multi-feature/multi-dimensional' (MF/MD) to describe his approach.
8 See section 2.6.1 for a description of how normalised frequencies are calculated, and 2.6.2 for a brief introduction to factor analysis.
9 The statistics of clustering are described very briefly in section 2.6.2; readers who wish to explore the clustering techniques used by corpus linguists are referred to Oakes (1998: 95–104), Gries (2009b: 306–19; 2010a), or Baayen (2008: 127–59).
10 We will return to the topic of intersections between corpus linguistics and functional linguistics in Chapter 7.
11 This sequence occurs ninety-one times in the BNC.
12 The extralinguistic variables in this case are metadata pertaining to the speakers in the corpus, such as age, ethnicity and so on.

6 Neo-Firthian corpus linguistics

1 Kennedy (1998: 14, 108) traces the basic idea back to the middle of the eighteenth century via the works of Alexander Cruden.
2 In this quotation from Stubbs, the notation *f(n,c)* should be read as 'the frequency of the node word and collocate occurring together in the corpus'.
3 Or related phenomena such as semantic prosody, which we will discuss later in this chapter.
4 SketchEngine is a good example of a system which incorporates syntactic structures into its support for analysing collocation – see section 2.5.4 for more discussion of this program.
5 See Stubbs (2001: 67–8) for a brief discussion of the variability of collocates across registers.
6 Data taken from the BE06 corpus (Baker 2009).
7 See Taylor (2008) for an interesting, corpus-based, exploration of different conceptual approaches to corpus linguistics.

7 Corpus methods and functionalist linguistics

1 As exemplified by Stefanowitsch and Gries (2003, 2005); Gries and Stefanowitsch (2004); and Gries and Divjak (2009); see section 7.5.
2 See http://spraakbanken.gu.se/parole/.
3 Two examples with the verb *like*, from the BNC: *The style of life, the quality of life I like very much in this country* [text CGB]; *'Some things I like to do quickly, Fuzz,' he said* [text G3G].

8 Corpus, psycholinguistics and functionalism

1 A *synapse* is a connection between two nerve cells (neurons).
2 This particular sequence occurs six times in the BNC.
3 To clarify: the transitive is *X (verb) Y*, the *to*-dative is *X (verb) Y to Z* and the ditransitive is *X (verb) Z Y*. The syntactic role of '(direct) object' can then be seen as an abstraction emphasising the shared properties of the slots labelled Y.

9 Conclusion

1 But see Mukherjee (2010) for an argument to the contrary.
2 This area is sometimes called *language generation* (see McDonald 2010).
3 For this reason, information extraction and information retrieval are major fields of research in computational linguistics (see Hobbs and Riloff 2010; Savoy and Gaussier 2010).
4 See www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05.
5 For an opposing argument, see Vepstas (2010).
6 See http://eebo.chadwyck.com.
7 It is a necessary prerequisite for these procedures that the machine-readable text of a database should be available and that the licence for its use should allow such processing; see Dunning *et al*. (2009) for a discussion of this issue.

References

Aarts, J. 1991. 'Intuition-based and observation-based grammars', in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*, pp. 44–63. London: Longman.

2002. 'Does corpus linguistics exist? Some old and new issues', in L. Brievik and A. Hasselgren (eds.) *From the COLT's Mouth: Language Corpora Studies, in Honour of Anna-Brita Stenström*, pp. 1–18. Amsterdam: Rodopi.

Abbès, R. and Dichy, J. 2008. 'AraConc, an Arabic concordance software based on the DIINAR.1 langauge resource', in *Proceedings of INFOS2008: The Sixth International Conference on Informatics and Systems-Special Track on Natural Language Processing, 27–28 March 2008*, pp. 127–34. Cairo: Cairo University.

Abercrombie, D. 1965. 'Pseudo-procedures in linguistics', in D. Abercrombie (ed.) *Studies in Phonetics and Linguistics*, pp. 114–19. Oxford University Press.

Adolphs, S. 2006. *Introducing Electronic Text Analysis*. London: Routledge.

2008. *Corpus and Context: Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins.

Aijmer, K. 2009. 'So er I just sort I dunno I think it's just because . . . : a corpus study of *I don't know* and *dunno* in learners' spoken English', in A. Jucker, D. Schreier and M. Hundt (eds.) *Corpora: Pragmatics and Discourse – Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*, pp. 151–68. Amsterdam: Rodopi.

(ed.) 2009. *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Aijmer, K. and Altenberg, B. (eds.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Aijmer, K. and Stenström, A.-B. 2004. *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins.

Aijmer, K., Altenberg, B. and Johansson, M. (eds.) 1996. *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund, March 1994. Lund University Press.

Alderson, J. C. 1996. 'Do corpora have a role in language assessment?', in J. Thomas and M. Short (eds.) *Using Corpora in Language Research*, pp. 248–59. London: Longman.

2007. 'Judging the frequency of English words', *Applied Linguistics* 28 (3): 383–409.

Allen, W. H., Beal, J. C., Corrigan, K. P., Maguire, W. and Moisl, H. L. 2007. 'A linguistic "time-capsule": the Newcastle Electronic Corpus of Tyneside English', in J. C. Beal, K. P. Corrigan and H. L. Moisl (eds.) *Creating and Digitising Language Corpora*, Vol. II: *Diachronic Databases*, pp. 16–48. Houndmills: Palgrave Macmillan.

Alsop, S. and Nesi, H. 2009. 'Issues in the development of the British Academic Written English (BAWE) corpus', *Corpora* 4 (1): 71–83.

Altenberg, B. 1989. Review of D. Biber (1988), *Variation across Speech and Writing. Studia Linguistica* 43 (2): 167–74.

Altenberg, B. and Aijmer, K. 2000. 'The English–Swedish Parallel Corpus: a resource for contrastive research and translation studies', in C. Mair and M. Hundt (eds.) *Corpus Linguistics and Linguistic Theory*, pp. 15–33. Amsterdam: Rodopi.

Amador-Moreno, C. P., O'Riordan, S. and Chambers, A. 2006. 'Integrating a corpus of classroom discourse in language teacher education: the case of discourse markers', *Recall* 18 (1): 83–104.

Anderson, W. and Corbett, J. 2009. *Exploring English with Online Corpora*. Basingstoke: Palgrave Macmillan.

Andor, J. 2004. 'The master and his performance: an interview with Noam Chomsky', *Intercultural Pragmatics* 1 (1): 93–112.

Anthony, L. 2005. 'AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit', in *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 7–13. Tokyo: Waseda University.

2009. 'Issues in the design and development of software tools for corpus studies: the case for collaboration', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 87–104. London: Continuum.

Archer, D. 2005. *Questions and Answers in the English Courtroom (1640–1760): A Sociopragmatic Analysis*. Amsterdam: John Benjamins.

Archer, D. and Culpeper, J. 2003. 'Sociopragmatic annotation: new directions and possibilities in historical corpus linguistics', in A. Wilson, P. Rayson and T. McEnery (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, pp. 37–58. Frankfurt am Main: Peter Lang.

Archer, D., McEnery, T., Rayson, P. and Hardie, A. 2003. 'Developing an automated semantic analysis system for Early Modern English', in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL Technical Papers Vol. 16, pp. 22–31. Lancaster University: UCREL.

Arnold, E. J., Wasow, T., Losongco, A. and Ginstrom, R. 2000. 'Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering', *Language* 76 (1): 28–55.

Ashby, W. J. 1981. 'The loss of the negative particle *ne* in French: a syntactic change in progress', *Language* 57 (3): 674–87.

Aston, G. and Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Atwell, E. S., Hughes, J. S. and Souter, C. 1994. 'AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models', in J. Klavans (eds.) *Proceedings of the ACL Workshop on The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pp. 21–8. Association for Computational Linguistics.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H. and Renouf, A. 1996. 'Chronicling the *Times*: productive lexical innovation in an English newspaper', *Language* 72 (1): 69–96.

Babarczy, A., Carroll, J. and Sampson, G. 2006. 'Definitional, personal, and mechanical constraints on part of speech annotation performance', *Natural Language Engineering* 12 (1): 77–90.

Bailey, C. 2010. 'Introduction: making knowledge visual', in C. Bailey and H. Gardiner (eds.) *Revisualizing Visual Culture*, pp. 1–11. Farnham: Ashgate.

Baker, M. 1993. 'Corpus linguistics and translation studies: implications and applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, pp. 233–352. Amsterdam: John Benjamins.

1995. 'Corpora in translation studies: an overview and some suggestions for future research', *Target* 7 (2): 223–43.

1999. 'The role of corpora in investigating the linguistic behaviour of professional translators', *International Journal of Corpus Linguistics* 4 (2): 281–98.

Baker, P. 1997. 'Consistency and accuracy in correcting automatically tagged corpora', in R. Garside, G. Leech and T. McEnery (eds.) *Corpus Annotation*, pp. 243–50. Harlow: Longman.

2004. '"Unnatural acts": discourses of homosexuality within the House of Lords debates on gay male law reform', *Journal of Sociolinguistics* 8 (1): 88–106.

2006. *Using Corpora in Discourse Analysis*. London: Continuum.

2008. *Sexed Texts*. London: Equinox.

2009. 'The BE06 corpus of British English and recent language change', *International Journal of Corpus Linguistics* 14 (3): 312–37.

2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press.

Baker, P., McEnery, T., Leisher, M., Cunningham H. and Gaizauskas, R. 2000. 'Mapping multiple South Asian 8 bit character sets to the Unicode standard', in *Proceedings of the Linguistic Exploration Workshop*, University of Pennsylvania. Available online at: www.ldc.upenn.edu/exploration/expl2000/papers/mcenery/mcenery.pdf.

Baker, P., Hardie, A., McEnery, T., Xiao, R. Z., Bontcheva, K., Cunningham, H., Gaizauskas, R., Hamza, O., Maynard, D., Tablan, V., Ursu, C., Jayaram, B. D. and Leisher, M. 2004. 'Corpus linguistics and South Asian languages: corpus creation and tool development', *Literary and Linguistic Computing* 19 (4): 509–24.

Baker, P., Hardie A. and McEnery, T. 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. 2008. 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse and Society* 19 (3): 273–306.

Ball, C. N. 1994. 'Automated text analysis: cautionary tales', *Literary and Linguistic Computing* 9 (4): 295–302.

Bandhu, C. M. 1971. *The Computer Concordance of Spoken Nepali*. Oklahoma: Summer Institute of Linguistics.

Barlow, M. 1995. *A Guide to ParaConc*. Houston, TX: Athelstan.

2000. *MonoConc Pro*. Houston, TX: Athelstan.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh University Press.

Baron, A., Rayson, P. and Archer, D. 2009. 'Word frequency and key word statistics in historical corpus linguistics', *Anglistik: International Journal of English Studies* 20 (1): 41–67.

Baroni, M. and Bernardini, S. 2004. 'BootCaT: Bootstrapping corpora and terms from the web', in *Proceedings of LREC 2004*, pp. 1313–16. Paris: European Language Resources Association (ELRA).

Baroni, M., Chantree, F., Kilgarriff, A. and Sharoff, S. 2008. 'CleanEval: a competition for cleaning webpages', in *Proceedings of LREC 2008*, pp. 638–43. Paris: European Language Resources Association (ELRA).

Barron, A. and Schneider, K. P. 2009. 'Variational pragmatics: studying the impact of social factors on language use in interaction', *Intercultural Pragmatics* 6 (4): 425–42.

Bauer, L. 1993. *Manual of Information to Accompany the Wellington Corpus of Written New Zealand English*. Wellington: Department of Linguistics, Victoria University. Available online at: http://khnt.hit.uib.no/icame/manuals/wellman/index.htm.

Beal, J. C., Corrigan, K. P. and Moisl, H. L. 2007. 'Taming digital voices and texts: models and methods for handling unconventional synchronic corpora', in J. C. Beal, K. P. Corrigan and H. Moisl (eds.) *Creating and Digitising Language Corpora*, Vol. I: *Synchronic Databases*, pp. 1–15. Houndmills: Palgrave Macmillan.

Berez, A. L. and Gries, St. Th. 2009. 'In defense of corpus-based methods: a behavioral profile analysis of polysemous *get* in English', in S. Moran, D. S. Tanner and M. Scanlon (eds.) *Proceedings of the 24th Northwest Linguistics Conference*. University of Washington Working Papers in Linguistics Vol. 27, pp. 157–66. Seattle, WA: Department of Linguistics.

Berg, T. and Abd-El-Jawad, H. 1996. 'The unfolding of suprasegmental representations: a cross-linguistic perspective', *Journal of Linguistics* 32 (2): 291–324.

Berger, A., Brown, P., Della Pietra, S., Della Pietra, V., Gillett, J., Lafferty, J., Mercer, R., Printz, H. and Ureš, L. 1994. 'The Candide system for machine translation', in *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 157–62. Advanced Research Projects Agency.

Berglund, Y. 2000. 'Utilising present-day English corpora: a case study concerning expressions of future', *ICAME Journal* 24: 25–64.

Berry-Rogghe, G. L. 1973. 'The computation of collocations and their relevance in lexical studies', in A. J. Aitken, R. Bailey and N. Hamilton-Smith (eds.) *The Computer and Literary Studies*, pp. 103–11. Edinburgh University Press.

Besnier, N. 1988. 'The linguistic relationships of spoken and written Nukulaelae registers', *Language* 64 (4): 707–36.

1998. Review of D. Biber (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison. Language in Society* 27 (1): 126–9.

Biber, D. 1986. 'Spoken and written textual dimensions in English: resolving the contradictory findings', *Language* 62 (2): 384–414.

1988. *Variation across Speech and Writing*. Cambridge University Press.

1989. 'A typology of English texts', *Linguistics* 27: 3–43.

1990. 'Methodological issues regarding corpus-based analyses of linguistic variation', *Literacy and Linguistic Computing* 5 (4): 257–69.

1993. 'Representativeness in corpus design', *Literary and Linguistic Computing* 8 (4): 243–57.

1995a. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge University Press.

1995b. 'On the role of computational, statistical, and interpretive techniques in multidimensional analyses of register variation: a reply to Watson', *Text* 15 (3): 341–70.

2004. 'Modal use across registers and time', in A. Curzan and K. Emmons (eds.) *Studies in the History of the English Language*, Vol. II: *Unfolding Conversations*, pp. 189–216. Berlin: Mouton de Gruyter.

2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
2009. 'A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing', *International Journal of Corpus Linguistics* 14 (3): 275–311.
Biber, D. and Conrad, S. 1999. 'Lexical bundles in conversation and academic prose', in H. Hasselgård and S. Oksefjell (eds.) *Out of Corpora: Studies in Honour of Stig Johansson*, pp. 181–90. Amsterdam: Rodopi.
2009. *Register, Genre and Style*. Cambridge University Press.
Biber, D. and Finegan, E. 1989. 'Drift and the evolution of English style: a history of three genres', *Language* 65 (3): 487–517.
Biber, D. and Jones, J. K. 2005. 'Merging corpus linguistic and discourse analytic research goals: discourse units in biology research articles', *Corpus Linguistics and Linguistic Theory* 1 (2): 151–82.
Biber, D., Finegan, E. and Atkinson, D. 1993. 'ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers', in J. Aarts, P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*, pp. 1–13. Amsterdam: Rodopi.
Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
Biber, D., Conrad, S. and Cortes, V. 2004. 'If you look at . . . : lexical bundles in university teaching and textbooks', *Applied Linguistics* 25 (3): 371–405.
Biber, D., Davies, M., Jones, J. K. and Tracy-Ventura, N. 2006. 'Spoken and written register variation in Spanish: a multi-dimensional analysis', *Corpora* 1 (1): 1–37.
Birner, B. J. 1994. 'Information status and word order: an analysis of English inversion', *Language* 70 (2): 233–59.
Black, E., Leech, G. and Garside, R. (eds.) 1993. *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach*. Amsterdam: Rodopi.
Bloom, L. 1973. *One Word at a Time: The Use of Single Word Utterances*. The Hague: Mouton.
Bloom, P. 1990. 'Subjectless sentences in child language', *Linguistic Inquiry* 21 (4): 491–504.
Boas, F. 1940. *Race, Language and Culture*. New York: Macmillan.
Bolinger, D. 1977. *Meaning and Form*. London: Longman.
Bolton, K., Nelson, G. and Hung, J. 2003. 'A corpus-based study of connectors in student writing: research from the International Corpus of English in Hong Kong (ICE-HK)', *International Journal of Corpus Linguistics* 7 (2): 165–82.
Borin, L. (ed.) 2001. *Parallel Corpora, Parallel Worlds*. Amsterdam: Rodopi.
Botley S. P. and McEnery, T. (eds.) 2000. *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam and Philadelphia: John Benjamins.
Boulton, A. 2009. 'Testing the limits of data driven learning: language proficiency and training', *ReCall* 21 (1): 37–51.
Brazil, D. 1995. *A Grammar of Speech*. Oxford University Press.
Breton, R. P., Kaspi, V. M., Kramer, M., McLaughlin, M. A., Lyutikov, M., Ransom, S. M., Stairs, I. H., Ferdman, R. D., Camilo, F. and Possenti, A. 2008. 'Relativistic spin precision in the double pulsar', *Science* 321 (5885): 104–7.

Brill, E. 1995. 'Transformation-based error-driven learning and Natural Language Processing: a case study in part-of-speech tagging', *Computational Linguistics* 21 (4): 543–65.

Brown, R. W. 1973. *A First Language: the Early Stages*. Cambridge, Massachusetts: Harvard University Press.

Brown, P., Lai, J. and Mercer, R. 1991. 'Aligning sentences in parallel corpora', in *Proceedings of the Twenty-Ninth Annual Meeting of the Association for Computational Linguistics*, pp. 169–76. Berkeley, CA.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. and Roossin, P. 1988. 'A statistical approach to language translation', in *Proceedings of COLING 1988*, pp. 71–6. Budapest.

Brugman, H., Russel, A., Broeder, D. and Wittenburg, P. 2002. 'EUDICO. Annotation and exploitation of multi media corpora over the Internet', in *Proceedings of the Workshop on Data Architectures and Software Support for Large Corpora*, pp. 6–9. Paris: European Language Resources Association (ELRA).

Burnard, L. 2005. 'Metadata for corpus work', in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 30–46. Oxford: Oxbow Books.

Burnard, L. and McEnery, T. (eds.) 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Berlin: Peter Lang.

Cahnmann, M., Rymes, B. and Souto-Manning, M. 2005. 'Using critical discourse analysis to understand and facilitate identification processes of bilingual adults becoming teachers', *Critical Inquiry in Language Studies* 2 (4): 195–213.

Calude, A. S. 2008. 'Clefting and extraposition in English', *ICAME Journal* 32: 7–34.

Cameron, D. 1995. *Verbal Hygiene*. London: Routledge.

Cameron-Faulkner, T., Lieven, E. and Tomasello, M. 2003. 'A construction-based analysis of child directed speech', *Cognitive Science* 27 (6): 843–73.

Carden, G. 1982. 'Backwards anaphora in discourse context', *Journal of Linguistics* 18 (2): 361–87.

Carter, R. 2004. *Language and Creativity: The Art of Common Talk*. London: Routledge.

Carter, R. and Adolphs, S. 2008. 'Linking the verbal and visual: new directions for Corpus Linguistics', in A. Gerbig and O. Mason (eds.) *Language, People, Numbers: Corpus Linguistics and Society*, pp. 275–91. Amsterdam: Rodopi.

Carter, R. and McCarthy, M. 1995. 'Grammar and the spoken language', *Applied Linguistics* 16 (2): 141–58.

1997. *Exploring Spoken English*. Cambridge University Press.

2001. 'Ten criteria for a spoken grammar', in E. Hinkel and S. Fotos (eds.) *New Perspectives on Grammar Teaching in Second Language Classrooms*, pp. 31–76. Hillsdale, NJ: Lawrence Erlbaum.

2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge University Press.

Chandler, B. 1989. *Longman Mini Concordancer*. Harlow: Longman.

Cheshire, J. 1982. *Variation in an English Dialect*. Cambridge University Press.

Chomsky, N. 1957. *Syntactic Structures*. The Hague and Paris: Mouton.

1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Christ, O. 1994. 'A modular and flexible architecture for an integrated corpus query System', in *Proceedings of COMPLEX'94*, pp. 23–32. Budapest.

Chuang, F.-Y. and Nesi, H. 2006. 'An analysis of formal errors in a corpus of Chinese student writing', *Corpora* 1 (2): 251–71.

Chung, S.-F. 2008. 'Cross-linguistic comparisons of the MARKET metaphors', *Corpus Linguistics and Linguistic Theory* 4 (2): 141–75.

Church, K. and Hanks, P. 1989. 'Word association norms, mutual information, and lexicography', in *Proceedings of the Twenty-Seventh Annual Meeting of the Association for Computational Linguistics*, pp. 76–83. Vancouver: University of British Columbia.

Claridge, C. 2007. 'Constructing a corpus from the web: message boards', in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, pp. 87–108. Amsterdam: Rodopi.

Clark, E. V. and Carpenter, K. L. 1989. 'The notion of source in language acquisition', *Language* 65 (1): 1–30.

Clear, J. 1993. 'From Firth principles: collocation tools for the study of collocation', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, pp. 271–92. Amsterdam: John Benjamins.

Cohen, K. B. 2010. 'BioNLP: biomedical text mining', in N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing* (second edition). Boca Raton, FL: CRC Press.

Collins, P. and Peters, P. 1988. 'The Australian corpus project', in M. Kytö, O. Ihalainen and M. Rissanen (eds.) *Corpus Linguistics Hard and Soft*, pp. 103–20. Amsterdam: Rodopi.

Coulthard, M. and Johnson, A. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. London: Routledge.

Crasborn, O. 2008. 'Open access to sign language corpora', in O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood and E. Thoutenhoofd (eds.) *Construction and Exploitation of Sign Language Corpora*, pp. 33–8. Third Workshop on the Representation and Processing of Sign Languages. Paris: European Language Resources Association (ELRA).

Croft, W. 2001. *Radical Construction Grammar:Syntactic Theory in Typological Perspective*. Oxford University Press.

Crossley, S. A. and Louwerse, M. 2007. 'Multi-dimensional register classification using bigrams', *International Journal of Corpus Linguistics* 12 (4): 453–78.

Crowdy, S. 1995. 'The BNC spoken corpus', in G. Leech, G. Myers and J. Thomas (eds.) *Spoken English on Computer: Transcription, Mark-up and Application*, pp. 224–35. Harlow: Longman.

Cruttenden, A. 1978. 'Assimilation in child language and elsewhere', *Journal of Child Language* 5 (2): 373–8.

Culpeper, J. 2009. 'Keyness: words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*', *International Journal of Corpus Linguistics* 14 (1): 29–59.

Culpeper, J. and Kytö, M. 2000. 'Data in historical pragmatics: spoken interaction (re)cast as writing', *Journal of Historical Pragmatics* 1 (2): 175–99.

2002. 'Lexical bundles in Early Modern English: a window into the speech-related language of the past', in T. Fanego, B. Méndez-Naya and E. Seoane (eds.) *Sounds, Words, Texts and Change: Selected papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000*, pp. 45–65. Amsterdam: John Benjamins.

2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge University Press.

Curran, J. 2004. 'From distributional to semantic similarity', unpublished PhD thesis, University of Edinburgh.

Dahlmann, I. and Adolphs, S. 2009. 'Multi-modal spoken corpus analysis and language description: the case of multi-word expressions', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 125–39. London: Continuum.

Daille, B., Gaussier, E. and Langé, J.-M. 1996. 'An evaluation of statistical scores for word association', in J. Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Lévy and E. Vallduví (eds.) *The Tbilisi Symposium on Logic, Language and Computation: Selected Papers*, pp. 179–88. Stanford, CA: CSLI Publications.

Davies, M. 2005. 'The advantage of using relational databases for large corpora: speed, advanced queries and unlimited annotation', *International Journal of Corpus Linguistics* 10 (3): 307–34.

2009a. 'Word frequency in context: alternative architectures for examining related words, register variation and historical change', in D. Archer (ed.) *What's in a Wordlist? Investigating Word Frequency and Keyword Extraction*, pp. 53–68. Farnham: Ashgate.

2009b. 'The 385+ million word corpus of contemporary American English (1990–2008+): design, architecture and linguistic insights', *International Journal of Corpus Linguistics* 14 (2): 159–90.

2010. 'More than a peephole: using large and diverse online corpora', *International Journal of Corpus Linguistics* 15 (3): 412–18.

Declerck, R. and Reed, S. 2000. 'The semantics and pragmatics of *unless*', *English Language and Linguistics* 4 (2): 205–41.

De Cock, S. 1998. 'A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English', *International Journal of Corpus Linguistics* 3 (1): 59–80.

Deegan, M. and Sutherland, K. (eds.) 2009. *Text Editing, Print and the Digital World*. Farnham: Ashgate.

Deignan, A. 1999a. 'Linguistic metaphors and collocation in non-literary corpus data', *Metaphor and Symbol* 14: 19–38.

1999b. 'Metaphorical polysemy and paradigmatic relations: a corpus study', *Word* 50: 319–38.

2005. *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Déprez, V. and Pierce, A. 1993. 'Negation and functional projections in early grammar', *Linguistic Inquiry* 24 (1): 25–67.

DeRose, S. 1988. 'Grammatical category disambiguation by statistical optimization', *Computational Linguistics* 14 (1): 31–9.

Di Sciullo, A.-M., Muysken, P. and Singh, R. 1986. 'Government and code-mixing', *Journal of Linguistics* 22 (1): 1–24.

Diani, G. 2008. 'Emphasizers in spoken and written academic discourse: The case of *really*', *International Journal of Corpus Linguistics* 13 (3): 296–321.

Diessel, H. and Tomasello, M. 2005. 'Particle placement in early child language: a multifactorial analysis', *Corpus Linguistics and Linguistic Theory* 1 (1): 89–111.

Dik, S. 1997. *The Theory of Functional Grammar*, Part 1: The Structure of the Clause (second edition, ed. Kees Hengeveld). Berlin: Mouton de Gruyter.

Divjak, D. 2006. 'Ways of intending: delineating and structuring near-synonyms', in St. Th. Gries and A. Stefanowitsch (eds.) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, pp. 19–56. Berlin and New York: Mouton de Gruyter

Divjak, D. S. and Gries, St. Th. 2006. 'Ways of trying in Russian: clustering behavioral profiles', *Corpus Linguistics and Linguistic Theory* 2 (1): 23–60.

Dixon, R. M. W. 1979. 'Ergativity', *Language* 55 (1): 59–138.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge University Press.

Dons, U. 2004. *Descriptive Adequacy of Early Modern English Grammars*. Berlin: Mouton de Gruyter.

Downing, P. 1993. 'Pragmatic and semantic constraints on numeral quantifier position in Japanese', *Journal of Linguistics* 29 (1): 65–93.

Doyle, P. 2005. 'Replicating corpus-based linguistics: investigating lexical networks in text', in *Proceedings from Corpus Linguistics 2005*. University of Birmingham. Available online at: www.corpus. bham.ac.uk/conference/proceedings.shml.

Du Bois, J. W. 1987. 'The discourse basis of ergativity', *Language* 63 (4): 805–55.

Dulay, H. and Burt, M. 1973. 'Should we teach children syntax?', *Language Learning* 23: 95–123.

Dunning, T. 1993. 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics* 19 (1): 61–74.

Dunning, A., Gregory, I. and Hardie, A. 2009. 'Freeing up digital content: new research means new licenses', *Serials* 22 (2): 166–73.

Ebeling, S. and Heuboeck, A. 2007. 'Encoding document information in a corpus of student writing: the British Academic Written English Corpus', *Corpora* 2 (2): 241–56.

Ellis, N. C. 1998. 'Emergentism, connectionism and language learning', *Language Learning* 48 (4): 631–64.

2002. 'Frequency effects in language acquisition: a review with implications for theories of implicit and explicit language acquisition', *Studies in Second Language Acquisition* 24 (2): 143–88.

2003. 'Constructions, chunking, and connectionism: the emergence of second language structure', in C. Doughty and M. H. Long (eds.) *Handbook of Second Language Acquisition*, pp. 33–68. Oxford: Blackwell.

Ellis, N. C. and Cadierno, T. 2009. 'Constructing a second language. Introduction to the special section', *Annual Review of Cognitive Linguistics* 7 (1): 111–39.

Ellis, N. C. and Frey, E. 2009. 'The psycholinguistic reality of collocation and semantic prosody (2): affective priming', in R. Corrigan, E. Moravcsik, H. Ouali and K. Wheatley (eds.) *Formulaic Language*, pp. 473–97. Amsterdam: John Benjamins.

Ellis, N. C. and Simpson-Vlach, R. C. 2009. 'Formulaic language in native speakers: triangulating psycholinguistics, corpus linguistics, and education', *Corpus Linguistics and Linguistic Theory* 5 (1): 61–78.

Ellis, N. C., Frey E. and Jalkanen, I. 2009. 'The psycholinguistic reality of collocation and semantic prosody (1): lexical access', in U. Römer and R. Schulze (eds.) *Exploring the Lexis–Grammar Interface*, pp. 89–114. Amsterdam: John Benjamins.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M. and Laporte, N. 2001. *Cronfa Electroneg o Gymraeg, a 1 million word lexical database and frequency count for Welsh*. Accessed January 2010, at: www.bangor.ac.uk/ar/cb/ceg.php.en.

Ellis, R. 2008. *The Study of Second Language Acquisition* (second edition). Oxford University Press.

Ensslin, A. and Johnson, S. 2006. 'Language in the news: investigations into representations of "Englishness" using *WordSmith Tools*', *Corpora* 1 (2): 153–85.

Esser, J. 1999. 'Collocation, colligation, semantic preference and semantic prosody: new developments in the study of syntagmatic word relations', in W. Falkner and H.-J. Schmid (eds.) *Words, Lexemes and Concepts: Approaches to The Lexicon*, pp. 155–66. Tübingen: Gunter Narr Verlag.

Evert, S. 2005. 'The statistics of word cooccurrences: word pairs and collocations', unpublished PhD thesis, University of Stuttgart.

2008. 'Corpora and collocations', in A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics: An International Handbook*, pp. 1212–48. Berlin: Mouton de Gruyter.

Faber, D. and Lauridsen, K. 1991. 'The compilation of a Danish–English–French corpus in contract law', in S. Johansson and A. Stenström (eds.) *English Computer Corpora: Selected Papers and Research Guide*, pp. 235–43. Berlin: Mouton de Gruyter.

Fairclough, N. 2000. *New Labour, New Language?* London: Routledge.

Feldman, J. and Narayanan, S. 2004. 'Embodied meaning in a neural theory of language', *Brain and Language* 89 (2): 385–392.

Feldman, R. and Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fillmore, C. 1992. '"Corpus linguistics" or "Computer-aided armchair linguistics"', in J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, pp. 35–60. Berlin: Mouton de Gruyter.

Firth, J. R. 1957. *Papers in Linguistics*. Oxford University Press.

1968. 'A synopsis of linguistic theory 1930–55', in F. R. Palmer (ed.) *Selected Papers of J.R. Firth 1952–59*, pp. 168–205. London: Longman.

Fligelstone, S. 1992. 'Developing a scheme for annotating text to show anaphoric relations', in G. Leitner (ed.) *New Directions in English Language Corpora: Methodology, Results, Software Developments*, pp. 153–70. Berlin: Mouton de Gruyter.

Flowerdew, L. 1997. 'Interpersonal strategies: investigating interlanguage corpora', *RELC Journal* 28 (1): 72–88.

Fodor, J. A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Fortanet, I. 2004. 'Verbal stance in spoken academic discourse', in G. Camiciotti and E. Tognini Bonelli (eds.) *Academic Discourse: New Insights into Evaluation*, pp. 99–120. Bern: Peter Lang.

Fox, B. A. 1987. 'The noun phrase accessibility hierarchy revisited: subject primacy or the absolutive hypothesis?', *Language* 63 (4): 856–70.

Francis, G. 1995. 'Corpus-driven grammar and its relevance to the learning of English in a cross-cultural situation', in A. Pakir (ed.) *English in Education: Multicultural Perspectives*. Singapore: Unipress.

Francis, G., Hunston, S. and Manning, E. 1996. *Collins Cobuild Grammar Patterns 1: Verbs*. London: Collins.

Francis, W. N. and Kučera, H. 1964. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers*. Providence, Rhode Island: Department of Linguistics, Brown University. Available online at: http://khnt.hit.uib.no/icame/manuals/brown/index.htm.

Frankenberg-Garcia, A. 2008. 'Suggesting rather special facts: a corpus-based study of distinctive lexical distributions in translated texts', *Corpora* 3 (2): 195–211.

Frankenberg-Garcia, A. and Santos, D. 2003. 'Introducing COMPARA, the Portuguese–English parallel corpus', in C. Zannettin, S. Bernardini and D. Stewart (eds.) *Corpora in Translator Education*, pp. 71–88. Manchester: St Jerome Publishing.

Fries, C. 1940. *American English Grammar*. New York: Appleton Century.

1952. *The Structure of English*. London: Longman.

Frisson, S. and Pickering, M. 2001. 'Obtaining a figurative interpretation of a word: support for underspecification', *Metaphor and Symbol* 16 (3/4): 149–71.

Fromont, R. and Hay, J. 2008. 'ONZE Miner: the development of a browser-based research tool', *Corpora* 3 (2): 173–93.

Gabrielatos, C. and Baker, P. 2008. 'Discourses of refugees and asylum seekers in the UK Press 1996–2007', *Journal of English Linguistics* 36 (1): 5–38.

Gabrielatos, C., Torgerson, E., Hoffmann, S. and Fox, S. 2010. 'A corpus-based socio-linguistic study of indefinite article forms in London English', *Journal of English Linguistics* 38 (1): 1–38.

Gale, W. and Church, K. 1993. 'A program for aligning sentences in bilingual corpora', *Computational Linguistics* 19 (1): 75–102.

Garman, M. 1990. *Psycholinguistics*. Cambridge University Press.

Garside, R. and McEnery, T. 1993. 'Treebanking: the compilation of a corpus of skeleton-parsed sentences', in E. Black, R. Garside and G. Leech (eds.) *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach*, pp. 17–35. Amsterdam: Rodopi.

Garside, R. and Smith, N. 1997. 'A hybrid grammatical tagger: CLAWS4', in R. Garside, G. Leech and T. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 102–21. London: Longman.

Garside, R., Leech, G. and McEnery, T. (eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Garside, R., Leech, G. and Sampson, G. 1987. *The Computational Analysis of English: A Corpus-based Approach*. Harlow: Longman.

Genetti, C. and Crain, L. D. 2003. 'Beyond preferred argument structure: sentences, pronouns, and given referents in Nepali', in J. W. Du Bois, L. E. Kumpf and W. J. Ashby (eds.) *Preferred Argument Structure: Grammar as Architecture for Function*. Amsterdam: John Benjamins.

Geschwind, N. 1974. *Selected Papers on Language and the Brain*. Dordrecht: Reidel.

Ghadessy, M. and Gao, Y. 2001. 'Small corpora and translation: comparing thematic organization in two languages', in M. Ghadessy, A. Henry and R. L. Roseberry (eds.) *Small Corpus Studies and ELT: Theory and Practice*, pp. 335–59. Amsterdam and Philadelphia: John Benjamins.

Giering, D., Graustein, G., Hoffmann, A., Kirsten, H., Neubert, A. and Thiele, W. 1979. *English Grammar: A University Handbook*. Leipzig: VEB Verlag Enzyklopädie.

Gilquin, G. 2006. 'The place of prototypicality in corpus linguistics: causation in the hot seat', in St. Th. Gries and A. Stefanowitsch (eds.) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, pp. 159–91. Berlin: Mouton de Gruyter.

Gilquin, G. and Gries, St. Th. 2009. 'Corpora and experimental methods: a state-of-the-art review', *Corpus Linguistics and Linguistic Theory* 5 (1): 1–26.

Gilquin G., Papp S. and Díez-Bedmar, M. B. (eds.) 2008. *Linking up Contrastive and Learner Corpus Research*. Amsterdam and Atlanta: Rodopi.

Givón, T. 1995. *Functionalism and Grammar*. Amsterdam: John Benjamins.

Gledhill, C. 2000. *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.

Goldacre, B. 2008. *Bad Science*. London: HarperCollins.

Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Granath, S. and Wherrity, M. 2005. 'Prepositions with *that*-clause complements in tagged corpora, with a special focus on *in that*', in *Proceedings from Corpus Linguistics 2005*. University of Birmingham. Available online at: www.corpus.bham.ac.uk/conference/proceedings.shtml.

Granger S. 1993a. 'The International Corpus of Learner English', in J. Aarts, P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*, pp. 57–69. Amsterdam: Rodopi.

1993b. 'The International Corpus of Learner English', *The European English Messenger* 2 (1): 34.

1994. 'The Learner Corpus: a revolution in applied linguistics', *English Today* 39 (10/3): 25–9.

1996. 'From CA to CIA and back: an integrated approach to computerised bilingual and learner corpora', in K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*, pp. 38–51. Lund, March 1994. Lund University Press.

1999. 'Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus', in H. Hasselgård and S. Oksefjell (eds.) *Out of Corpora: Studies in Honour of Stig Johansson*, pp. 191–202. Amsterdam and Atlanta: Rodopi.

2003. 'Error-tagged learner corpora and CALL: a promising synergy', *CALICO* (special issue on Error Analysis and Error Correction in Computer-Assisted Language Learning) 20 (3): 465–80.

Granger, S. and Meunier, F. (eds.) 2009. *Phraseology: An Interdisciplinary Perspective*. Amsterdam and Philadelphia: John Benjamins.

Granger, S. and Rayson, P. 1998. 'Automatic profiling of learner texts', in S. Granger (ed.) *Learner English on Computer*, pp. 119–31. London: Longman.

Granger S., Hung J. and Petch-Tyson S. (eds.) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam and Philadelphia: John Benjamins.

Granger, S., Meunier, F. and Paquot, M. 2009. *International Corpus of Learner English Version 2*. Louvain: Presses Universitaires de Louvain.

Greenbaum, S. 1974. 'Some verb-intensifier collocations in American and British English', *American Speech* 49 (1/2): 79–89.

Greenbaum, S. (ed.) 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Greenberg, J. H. 1963. 'Some universals of grammar with particular reference to the order of meaningful elements', in J. H. Greenberg (ed.) *Universals of Language*, pp. 73–113. Cambridge, MA: MIT Press.

Greene, B. B. and Rubin, G. M. 1971. *Automatic Grammatical Tagging of English*. Technical Report. Providence, RI: Department of Linguistics, Brown University.

Grefenstette, G. 1992. 'Use of syntactic context to produce term association lists for text retrieval', *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval*, pp. 89–97. New York: ACM.

Gregory, I. and Hardie, A. 2011. 'Visual GISting: bringing corpus linguistics and geographical information systems together'. *Literary and Linguistic Computing*.

Gries, St. Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York: Continuum.

2006a. 'Introduction', in St. Th. Gries and A. Stefanowitsch (eds.) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, pp. 1–17. Berlin and New York: Mouton de Gruyter.

2006b. 'Corpus-based methods and cognitive semantics: the many meanings of *to run*', in St. Th. Gries and A. Stefanowitsch (eds.) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, pp. 57–99. Berlin and New York: Mouton de Gruyter.

2006c. 'Exploring variability within and between corpora: some methodological considerations', *Corpora* 1 (2): 109–51.

2008. 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics* 13 (4): 403–37.

2009a. *Quantitative Corpus Linguistics with R*. London and New York: Routledge.

2009b. *Statistics for Linguistics with R: A Practical Introduction*. Berlin: Mouton de Gruyter.

2010a. 'Useful statistics for corpus linguistics', in A. Sánchez and M. Almela (eds.) *A Mosaic of Corpus Linguistics: Selected Approaches*, pp. 269–91. Frankfurt am Main: Peter Lang.

2010b. 'Corpus linguistics and theoretical linguistics: a love–hate relationship? Not necessarily . . . ', *International Journal of Corpus Linguistics* 15 (3): 327–43.

2010c. 'Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods', in T. Harris & M. Moreno Jaén (eds.) *Corpus Linguistics in Language Teaching*. Frankfurt am Main: Peter Lang.

Gries, St. Th. and Divjak, D. S. 2009. 'Behavioral profiles: a corpus-based approach towards cognitive semantic analysis', in V. Evans and S. S. Pourcel (eds.) *New Directions in Cognitive Linguistics*, pp. 57–75. Amsterdam and Philadelphia: John Benjamins.

Gries, St. Th. and Otani, N. 2010. 'Behavioral profiles: a corpus-based perspective on synonymy and antonymy', *ICAME Journal* 34: 121–50.

Gries, St. Th. and Stefanowitsch, A. 2004. 'Extending collostructional analysis: a corpus-based perspective on "alternations"', *International Journal of Corpus Linguistics* 9 (1): 97–129.

Gries, St. Th., Hampe, B. and Schönefeld, D. 2005. 'Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions', *Cognitive Linguistics* 16 (4): 635–76.

2010. 'Converging evidence II: more on the association of verbs and constructions', in J. Newman and S. Rice (eds.) *Empirical and Experimental Methods in Cognitive/Functional Research*, pp. 59–72. Stanford, CA: CSLI Publications.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R. and Wilson, R. 1989. 'The learnability and acquisition of the dative', *Language* 65 (2): 203–57.

Halliday, M. A. K. 1985. *Introduction to Functional Grammar*. London: Edward Arnold.
Hanks, P. 1996. 'Contextual dependency and lexical sets', *International Journal of Corpus Linguistics* 1 (1): 75–98.
2009. 'The impact of corpora on dictionaries', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 214–36. London: Continuum.
Hård af Segerstad, Y. 2005. 'Language in SMS – a socio-linguistic view', in R. Harper, L. Palen and A. Taylor (eds.) *The Inside Text: Social, Cultural and Design Perspectives on SMS*, pp. 33–51. Dordrecht: Kluwer Academic Publishers.
Hardie, A. 2004. 'The computational analysis of morphosyntactic categories in Urdu', unpublished PhD thesis, Lancaster University. Available online at http://eprints.lancs.ac.uk/106/.
2005. 'Automated part-of-speech analysis of Urdu: conceptual and technical issues', in Y. Yadava, G. Bhattarai, R. R. Lohani, B. Prasain and K. Parajuli (eds.) *Contemporary Issues in Nepalese Linguistics*, pp. 49–72. Kathmandu: Linguistic Society of Nepal.
2007. 'From legacy encodings to Unicode: the graphical and logical principles in the scripts of South Asia', *Language Resources and Evaluation* 41 (1): 1–25.
Forthcoming. 'CQPweb – combining power, flexibility and usability in a corpus analysis tool'.
Hardie, A. and McEnery, T. 2003. 'The *were*-subjunctive in British rural dialects: marrying corpus and questionnaire data', *Computers and the Humanities* 37 (2): 205–28.
2009. 'Corpus linguistics and historical contexts: text reuse and the expression of bias in early modern English journalism', in R. Bowen, M. Mobärg and S. Ohlander (eds.) *Corpora and Discourse – and Stuff: Papers in Honour of Karin Aijmer*, pp. 59–92. Göteborg: Acta Universitatis Gothoburgensis.
Hardie, A. and Mudraya, O. 2009. 'Collocational patterning in cross-linguistic perspective: adpositions in English, Nepali, and Russian', *Arena Romanistica* 4: 138–49.
Hardie, A., Baker, P., McEnery, T. and Jayaram, B. D. 2006. 'Corpus-building for South Asian languages', in A. Saxena and L. Borin (eds.) *Lesser Known Languages of South Asia*. Berlin: Mouton de Gruyter.
Hardie, A., Lohani, R. R., Regmi, B. R. and Yadava, Y. P. 2009. 'A morphosyntactic categorisation scheme for the automated analysis of Nepali', in R. Singh (ed.), *Annual Review of South Asian Languages and Linguistics 2009*, pp. 171–98. Berlin: Mouton de Gruyter.
Hardt-Mautner, G. 1995. 'How does one become a good European: the British press and European integration', *Discourse and Society* 6 (2): 177–205.
2000. *Der britische Europa-Diskurs: Methodenreflexion und Fallstudien zur Berichterstattung in der Tagespresse*. Wien: Passagen-Verlag.
Harris, A. 2006. 'Revisiting anaphoric islands', *Language* 82 (1): 114–30.
Hasund, K. 1998. 'Protecting the innocent: the issue of informants' anonymity in the COLT corpus', in A. Renouf (ed.) *Explorations in Corpus Linguistics*, pp. 13–28. Amsterdam: Rodopi.
Heath, S. 2010. 'Diversity and reuse of digital resources for ancient Mediterranean material culture', in G. Bodard and S. Mahony (eds.) *Digital Research in the Study of Classical Antiquity*, pp. 35–52. Farnham: Ashgate.
Heffer, C. and Sauntson, H. (eds.) 2000. *Words in Context: A Tribute to John Sinclair on His Retirement*. Birmingham: University of Birmingham.

Hindle, D. 1983. *User Manual for Fidditch*. Technical Memorandum 7590–142. USA: Naval Research Laboratory.

Hockey, S. 1988. *Micro-OCP (OCP Version 2)*. Oxford University Press.

Hobbs, J. R. and Riloff, E. 2010. 'Information extraction', in N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing* (second edition). Boca Raton, FL: CRC Press.

Hoey, M. 1997. 'From concordance to text structure: new uses for computer corpora', in B. Lewandowska and P. Melia (eds.) *Proceedings of the Practical Applications of Language Corpora Conference*, pp. 2–23. Łódź University Press.

2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

2007a. 'Lexical priming and literary creativity', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora: Theory and Analysis*, pp. 7–30. London: Continuum.

2007b. 'Grammatical creativity: a corpus perspective', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora: Theory and Analysis*, pp. 31–56. London: Continuum.

Hoey, M. and O'Donnell, M. 2008. 'Lexicography, grammar and textual position', *International Journal of Lexicography* 21 (3): 293–309.

Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W. (eds.) 2007. *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum.

Hoffmann, S. 2007a. 'From web page to mega-corpus: the CNN transcripts', in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, pp. 69–85. Amsterdam: Rodopi.

2007b. 'Processing internet-derived text – creating a corpus of usenet messages', *Literary and Linguistic Computing* 22 (2): 151–65.

Hoffmann, S. and Lehmann, H. M. 2000. 'Collocational evidence from the British National Corpus', in J. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, pp. 17–32. Amsterdam: Rodopi.

Hoffmann, S. and Mukherjee, J. 2007. 'Ditransitive verbs in Indian English and British English: a corpus-linguistic study', *Arbeiten aus Anglistik und Amerikanistik* 32 (1): 5–24.

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund Prytz, Y. 2008. *Corpus Linguistics with BNCweb: A Practical Guide*. Frankfurt am Main: Peter Lang.

Hofland. K. and Johansson, S. 1982. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.

Hollmann, W. B. 2005. 'Passivisability of English periphrastic causatives', in St. Th. Gries and A. Stefanowitsch (eds.) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, pp. 193–223. Berlin and New York: Mouton de Gruyter.

Hollmann, W. B. 2007. 'From language-specific constraints to implicational universals: a cognitive-typological view of the dative alternation', *Functions of Language* 14 (1): 57–78

Hollmann, W. B. and Siewierska, A. 2006. 'Corpora and (the need for) other methods in a study of Lancashire dialect', *Zeitschrift für Anglistik und Amerikanistik* 54 (2): 203–16.

Holmes, J. 1996. 'Collecting the Wellington Corpus of Spoken New Zealand English: some methodological challenges', *New Zealand English Journal* 10 (1): 10–15.

Holmes, J., Vine, B. and Johnson, G. 1998. *Guide to The Wellington Corpus of Spoken New Zealand English*. School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Hopper, P. J. and Traugott, E. C. 1993. *Grammaticalization*. Cambridge University Press.

Huang, C.-R., Chen, K.-J., Yang, Y.-Y. 1994. 'Character-based collocation for Mandarin Chinese', in *Proceedings of COLING-94*, pp. 540–3. Kyoto.

Hudson, R. 1994. 'About 37% of word-tokens are nouns', *Language* 70 (2): 331–9.

Hughes, G. 1998. *Swearing: A Social History of Foul Language, Oaths and Profanity in English* (second edition). London: Blackwell.

Hundt, M. 1998. 'It is important that this study (*should*) *be* based on the analysis of parallel corpora: on the use of the mandative subjunctive in four major varieties of English', in H. Lindquist, S. Klintborg, M. Levin and M. Estling (eds.) *The Major Varieties of English: Paper from MAVEN 97*, pp. 159–75. Acta Wexionensia Humaniora, Växjö University.

2004a. 'The passival and the progressive passive: a case study of layering in the English aspect and voice systems', in H. Lindquist and C. Mair (eds.) *Corpus Approaches to Grammaticalisation in English*, pp. 79–120. Amsterdam and Philadelphia: John Benjamins.

2004b. 'Animacy, agency and the spread of the progressive in modern English', *English Language and Linguistics* 8 (1): 47–69.

2007. *English Mediopassive Constructions: A Cognitive, Corpus-based Study of Their Origin, Spread and Current Status*. Amsterdam: Rodopi.

Hundt, M., Nesselhauf, N. and Biewer, C. 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Hundt, M., Sand, A. and Siemund, R. 1998. *Manual of Information to Accompany The Freiburg–LOB Corpus of British English ('FLOB')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg. Available online at: http://khnt.hit.uib.no/icame/manuals/flob/index.htm.

Hundt, M., Sand, A. and Skandera, P. 1999. *Manual of Information to Accompany the Freiburg–Brown Corpus of American English ('Frown')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg. Available online at: http://khnt.hit.uib.no/icame/manuals/frown/index.htm.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge University Press.

2007. 'Semantic prosody revisited', *International Journal of Corpus Linguistics* 12 (2): 249–68.

Hunston, S. and Francis, G. 1999. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Hunston, S. and Thompson, G. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford University Press.

Hunston, S, Francis, G. and Manning, E. 1998. *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: Collins.

Hyams, N. and Wexler, K. 1993. 'On the grammatical basis of null subjects in child language', *Linguistic Inquiry* 24 (3): 421–59.

Ide, N. and Reppen, R. 2004. 'The American National Corpus: overall goals and first release', *Journal of English Linguistics* 32 (2): 105–13.

Ide, N., Baker, C., Fellbaum, C. and Passonneau, R. 2010. 'MASC: a community resource for and by the people', in *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*, pp. 68–73. Uppsala, Sweden.

Inaki, A. and Okita, T. 2006. 'A small-corpus-based approach to Alice's roles', *Literary and Linguistic Computing* 25 (3): 283–94.

Ingram, J. C. L. 2007. *Neurolinguistics*. Cambridge: Cambridge University Press.

Izquierdo, M., Hofland, K. and Reigem, Ø. 2008. 'The ACTRES parallel corpus: an English–Spanish translation corpus', *Corpora* 3 (1): 31–41.

Johansson, S. 1998. 'On the role of corpora in cross-linguistic research', in S. Johansson and S. Oksefjell (eds.) *Corpora and Cross-linguistic Research: Theory, Method and Case Studies*, pp. 3–24. Amsterdam: Rodopi.

2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.

Johansson, S. and Hofland, K. 1994. 'Towards an English–Norwegian parallel corpus', in U. Fries, G. Tottie and P. Schneider (eds.) *Creating and Using English Language Corpora*, pp. 25–37. Amsterdam: Rodopi.

Johansson, S., Leech, G. and Goodluck, H. 1978. *Manual of Information to Accompany the Lancaster–Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo. Available online at: http://khnt.hit.uib.no/icame/manuals/lob/index.htm.

Johns, T. 1994. 'From printout to handout: grammar and vocabulary teaching in the context of Data-driven Learning', in T. Odlin (ed.) *Perspectives on Pedagogical Grammar*, pp. 293–313. Cambridge University Press.

Johns, T. 1997. 'Contexts: the background, development and trialling of a concordance-based CALL program', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) *Teaching and Language Corpora*, pp. 100–15. London: Longman.

Johnston, T. and Schembri, A. 2006. 'Issues in the creation is a digital archive of a signed language', in L. Barwick and N. Thieburger (eds.) *Sustainable Data from Digital Fieldwork*, pp. 7–16. University of Sydney Press.

Jones, R. 1997. 'Creating and using a corpus of spoken German', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) *Teaching and Language Corpora*, pp. 146–56. London: Longman.

Jones, R. and Tschirmer, E. 2006. *A Frequency Dictionary of German: Core Vocabulary for Learners*. London: Routledge.

Jones, S., Paradis, C., Murphy, L. M. and Wilners, C. 2007. 'Googling for opposites: a web-based study of antonym canonicity', *Corpora* 2 (2): 129–55.

Jucker, A., Fritz, G. and Lebsanft, F. 1999. *Historical Dialogue Analysis*. Amsterdam: John Benjamins.

Juilland, A. and Chang-Rodriguez, E. 1964. *A Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Jurafsky, D. 2003. 'Probabilistic modeling in psycholinguistics: linguistic comprehension and production', in R. Bod, J. Hay and S. Jannedy (eds.) *Probabilistic Linguistics*, pp. 39–95. Cambridge, MA: MIT Press.

Kachru, B. 1986. *The Alchemy of English: The Spread, Functions and Models of Non-native Englishes*. Oxford: Pergamon.

Kaleta, A. 2009. 'English aspectual verbs and their complements: the case of "begin"', in B. Lewandowska-Tomaszczyk and K. Dziwirek (eds.) *Studies in Cognitive Corpus Linguistics*, pp. 173–90. Frankfurt am Main: Peter Lang.

Kaltenböck, G. 2003. 'On the syntactic and semantic status of anticipatory *it*', *English Language and Linguistics* 7 (2): 235–55.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (eds.) 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Kaye, G. 1990. 'A corpus-builder and real time concordance browser for an IBM PC', in J. Aarts and W. Meijs (eds.) *Theory and Practice in Corpus Linguistics*, pp. 137–62. Amsterdam: Rodopi.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Harlow: Longman.

Kerswill, P. 1987. 'Levels of linguistic variation in Durham', *Journal of Linguistics* 23 (1): 25–49.

1993. 'Rural dialect speakers in an urban speech community: the role of dialect contact in defining a sociolinguistic concept', *International Journal of Applied Linguistics* 3 (1): 33–56.

Kerswill, P. and Williams, A. 2000. 'Creating a new town koine: children and language change in Milton Keynes', *Language in Society* 29 (1): 65–115.

Kettemann, B. and Mark, G. 2002. *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi.

KhosraviNik, M. 2009. 'The representation of refugees, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005)', *Discourse and Society* 20 (4): 477–98.

Kilgarriff, A. 1995. Review of D. Biber (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison. Journal of Natural Language Engineering* 1 (4): 611–13.

2005. 'Language is never, ever, ever, random', *Corpus Linguistics and Linguistic Theory* 1 (2): 263–76.

Kilgarriff, A. and Grefenstette, G. 2003. 'Introduction to the special issue on the Web as Corpus', *Computational Linguistics* 29 (3): 333–47.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. 'The Sketch Engine', in G. Williams and S. Vessier (eds.) *Proceedings of the Eleventh International Congress of Euralex 2004*, pp. 105–16. Bretagne, France: Université de Bretagne-Sud.

King, B. 2009. 'Building and analysing corpora of computer mediated communication', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 301–20. London: Continuum.

Kitis, E. 2009. 'Emotions as discursive constructs: the case of the psych-verb "fear"', in B. Lewandowska-Tomaszczyk and K. Dziwirek (eds.) *Studies in Cognitive Corpus Linguistics*, pp. 147–72. Frankfurt am Main: Peter Lang.

Knight, D., Evans, D., Carter, R. and Adolphs, S. 2009. 'HeadTalk, HandTalk and the corpus: towards a framework for multi-modal, multi-media corpus development', *Corpora* 4 (1): 1–32.

Knowles, G. 1993. 'From text to waveform: converting the Lancaster/IBM spoken English corpus into a speech database', in C. Souter and E. Atwell (eds.) *Corpus-based Computational Linguistics*, pp. 47–58. Amsterdam: Rodopi.

Knowles, G., Williams, B. and Taylor, L. 1996. *A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English Corpus*. London: Longman.

Koeva, S., Maurel, D. and Silberztein, M. (eds.) 2007. *Formaliser les langues avec l'ordinateur : de INTEX à NooJ*, Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté.

Koller, V. and Mautner, G. 2004. 'Computer applications in critical discourse analysis', in C. Coffin, A. Hewings and K. O'Halloran (eds.) *Applying English Grammar: Corpus and Functional Approaches*, pp. 216–28. London: Arnold.

Koller, V., Hardie, A., Rayson, P. and Semino, E. 2008. 'Using a semantic annotation tool for the analysis of metaphor in discourse', *Metaphorik.de* 15. Available online at: www.metaphorik.de/15/.

König, E. and Siemund, P. 2000. 'Logically free *self*-forms, logophoricity, and intensification in English', *English Language and Linguistics* 4 (2): 183–204.

Kovàcs, A. and Wodak, R. (eds.) 2003. *NATO, Neutrality and National Identity*. Vienna: Böhlau.

Krashen, S., Butler, J., Birnbaum, R. and Robertson, J. 1978. 'Two studies in language acquisition and language learning', *ITL Review of Applied Linguistics* 39/40: 73–92.

Krenn, B. and Evert, S. 2001. 'Can we do better than frequency? A case study on extracting PP-verb collocations', in *Proceedings of the ACL Workshop on Collocations*, pp. 39–46. Toulouse, France.

Krishnamurthy, R. 1996. 'Ethnic, racial and tribal: the language of racism?', in C. R. Caldas-Coulthard and M. Coulthard (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*, pp. 129–49. London: Routledge.

2000. 'Collocation: from *silly ass* to lexical sets', in C. Heffer and H. Sauntson (eds.) *Words in Context: A Tribute to John Sinclair on His Retirement*, pp. 31–47. University of Birmingham.

Kučera, H. and Francis, W. N. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Kytö, M. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts* (third edition). Department of English, University of Helsinki.

1997. 'BE/HAVE + past participle: the choice of the auxiliary with intransitives from late middle to modern English', in M. Rissanen, M. Kytö and K. Heikkonen (eds.) *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles*, pp. 17–85. Berlin and New York: Mouton de Gruyter.

Kytö, M. and Rissanen, M. 1992. 'A language in transition: the Helsinki Corpus of English Texts', *ICAME Journal* 16: 7–28.

Kytö, M. and Walker, T. 2006. *Guide to A Corpus of English Dialogues 1560–1760*. Uppsala: Acta Universitatis Upsaliensis.

Kytö, M., Rissanen, M. and Wright, S. (eds.) 1994. *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge*, 25–27 March 1993. Amsterdam and Atlanta, GA: Rodopi.

Labov, W. 1969. 'Contraction, deletion, and inherent variability of the English copula', *Language* 45 (4): 715–62.

1972. *Language in the Inner City: Studies in Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. University of Chicago Press.

Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By*. University of Chicago Press.
Langacker, R. W. 1987. *Foundations of Cognitive Grammar*, Vol. I: *Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
1991. *Foundations of Cognitive Grammar*, Vol. II: *Descriptive Application*. Stanford: Stanford University Press.
2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press.
Laviosa, S. 2002. *Corpus-based Translation Studies: Theory, Findings, Application*. Amsterdam: Rodopi.
Le Page, R. 1980. 'Theoretical aspects of sociolinguistic studies in pidgin and creole languages', in A. Valdman and A. Highfield (eds.) *Theoretical Orientations in Creole Studies*, pp. 331–51. New York: Academic Press.
Le Page, R. and Tabouret-Keller, A. 1985. *Acts of Identity*. Cambridge University Press.
Lee, D. Y. W. 2001. 'Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle', *Language Learning and Technology* 5 (3): 37–72. Available online at: http://llt.msu.edu/vol5num3/lee/default.html.
Leech, G. 1971. *Meaning and the English Verb*. London: Longman.
1992. 'Corpora and theories of linguistic performance', in J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, pp. 105–22. Berlin: Mouton de Gruyter.
1998. 'The special grammar of conversation', *Longman Language Review* 5: 9–14.
2004a. *Meaning and the English Verb* (third edition). Harlow: Pearson Education.
2004b. 'Recent grammatical change in English: data, description, theory', in K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics: Papers from the Twenty-Third International Conference on English Language Research on Computerized Corpora (ICAME 23)*, pp. 61–81. Göteborg, 22–26 May 2002. Amsterdam: Rodopi.
2007. 'New resources, or just better old ones?', in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, pp. 134–49. Amsterdam: Rodopi.
Leech, G. and Fallon, R. 1992. 'Computer corpora: what do they tell us about culture?', *ICAME Journal* 16: 29–50.
Leech, G. and Johansson, S. 2009. 'The coming of ICAME', *ICAME Journal* 33: 5–20.
Leech, G. and Smith, N. 2005. 'Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB', *ICAME Journal* 29: 83–98.
2006. 'Recent grammatical change in written English 1961–1992: some preliminary findings of a comparison of American with British English', in A. Renouf and A. Kehoe (eds.), *The Changing Face of Corpus Linguistics*, pp. 186–204. Amsterdam: Rodopi.
Leech, G. and Weisser, M. 2003. 'Generic speech act annotation for task-oriented dialogues', in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL Technical Papers, vol. 16, pp. 441–6. Lancaster University: UCREL.
Leech, G. and Wilson, A. 1994. 'Recommendations for the morphosyntactic annotation of corpora', *EAGLES Document EAG-TCWG-MAC/R*. Available online at:www.ilc.cnr.it/EAGLES/browse.html.
1999. 'Standards for tagsets', in H. van Halteren (ed.) *Syntactic Wordclass Tagging*, pp. 55–80. Dordrecht: Kluwer Academic Publishers.

Leech, G., Barnett, R. and Kahrel, P. 1995. 'Recommendations for the syntactic annotation of corpora', *EAGLES Document EAG-TCWG-SASG /*1.8. Available online at: www.ilc.cnr.it/EAGLES/browse.html.

Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge University Press.

Leitner, G. 1992. *New Directions in English Language Corpora: Methodology, Results, Software Development*. Berlin: Mouton de Gruyter.

Léon, J. 2005. 'Claimed and unclaimed sources of corpus linguistics', *Henry Sweet Society Bulletin* 44: 36–50.

Levis, J. and Cortes, V. 2008. 'Minimal pairs in spoken corpora: implications for pronunciation assessment and teaching', in C. A. Chapelle, Y.-R. Chung and J. Xu (eds.) *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*, pp. 197–208. Ames, IA: Iowa State University.

Lew, R. 2009. 'The web as corpus versus traditional corpora', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 289–300. London: Continuum.

Li, C. N. and Thompson, S. A. 1975. 'The semantic function of word order in Chinese', in C. N. Li (ed.) *Word Order and Word Order Change*, pp. 163–95. Austin: University of Texas Press.

Li, P., Farkas, I. and MacWhinney, B. 2004. 'Early lexical development in a self-organizing neural network', *Neural Networks* 17 (8/9): 1345–62.

Lin, D. 1998. 'Automatic retrieval and clustering of similar words', in *Proceedings of COLING-ACL 1998*, pp. 768–74. Montreal: Université de Montréal.

Liu, B. 2010. 'Sentiment analysis and subjectivity', in N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing* (second edition), pp. 626–66. Boca Raton, FL: CRC Press.

Lönngren, L. 1993. *Chastotnyj slovar' sovremennogo russkogo jazyka*. (A Frequency Dictionary of Modern Russian.) Studia Slavica Upsaliensia 32. Uppsala: Uppsala Universitet.

Louw, W. E. 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, pp. 157–76. Amsterdam: John Benjamins.

2000. 'Contextual prosodic theory: bringing semantic prosodies to life', in C. Heffer and H. Sauntson (eds.) *Words in Context: A Tribute to John Sinclair on his Retirement*, pp. 48–94. University of Birmingham.

2008. 'Consolidating empirical method in data-assisted stylistics: towards a corpus-attested glossary of literary terms', in S. Zyngier, M. Bortolussi, A. Chesnokova and J. Auracher (eds.) *Directions in Empirical Literary Studies: In Honour of Willie van Peer*, pp. 243–64. Amsterdam: John Benjamins.

2010. 'Collocation as instrumentation for meaning: a scientific fact', in W. van Peer, S. Zyngier and V. Viana (eds.) *Literary Education and Digital Learning: Methods and Technologies*, pp. 79–101. Hershey, PA: IGI Global.

Lüdeling, A. and Kytö, M. 2008. *Corpus Linguistics: An international Handbook*. Berlin: Mouton de Gruyter.

Luk, R. W. P. 1994. 'An IBM-PC environment for Chinese corpus analysis', *Proceedings of the 15th Conference in Computational Linguistics*, Vol. I, pp. 584–7. Morristown, NJ: Association for Computational Linguistics.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge University Press.

McCarthy, M. and Carter, R. 2001. 'Ten criteria for a spoken grammar', in E. Hinkel and S. Fotos (eds.) *New Perspectives on Grammar Teaching in Second Language Classrooms*, pp. 51–75. Mahwah, NJ: Lawrence Erlbaum.

McCarty, W. 2005. *Humanities Computing*. Basingstoke: Palgrave Macmillan.

Maclagan, M., Davis, B. and Lunsford, R. 2008. 'Fixed expressions, extenders and metonymy in the speech of people with Alzheimer's disease', in S. Granger and F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*, pp. 175–90. Amsterdam: John Benjamins.

McDonald, D. M. 2010. 'Natural language generation', in N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing* (second edition), pp. 121–46. Boca Raton, FL: CRC Press.

McDonald, S. A. and Shillcock, R. C. 2003a. 'Eye movements reveal the on-line computation of lexical probabilities during reading', *Psychological Science* 14 (6): 648–52.

2003b. 'Low-level predictive inference in reading: the influence of transitional probabilities on eye movements', *Vision Research* 43 (16): 1735–51.

McEnery, T. 2005. *Swearing in English: Bad Language, Purity, and Power from 1586 to the Present*. London: Routledge.

McEnery, T. and Kifle, N. 2001. 'Epistemic modality in the argumentative essays of second language writers', in J. Flowerdew (ed.) *Academic Discourse*, pp. 182–95. London: Longman.

McEnery, T. and Oakes, M. 1995. 'Sentence and word alignment in the Crater project: methods and assessment', in S. Warwick-Armstrong (ed.) *Proceedings of the Association for Computational Linguistics Workshop SIG-DAT Workshop*, pp. 104–16. Dublin: ACL.

McEnery, T. and Ostler, N. 2000. 'A new agenda for corpus linguistics: working with all of the world's languages', *Literary and Linguistic Computing* 15 (4): 403–30.

McEnery, T. and Wilson, A. 2001 (second edition). *Corpus Linguistics*. Edinburgh University Press.

McEnery, T. and Xiao, R. Z. 2004a. 'The Lancaster corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study', in M. Lino, M. Xavier, F. Ferreire, R. Costa and R. Silva (eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, pp. 1175–8. 24–30 May 2004. Lisbon.

2004b. 'Swearing in modern British English: the case of fuck in the BNC', *Language and Literature* 13 (3): 235–68.

2005a. 'Passive constructions in English and Chinese: a corpus-based contrastive study', in *Proceedings from Corpus Linguistics 2005*. University of Birmingham. Available at: www.corpus.bham.ac.uk/conference/proceedings.shtml.

2005b. '*Help* or *help to*: what do corpora have to say?', *English Studies* 86 (2): 161–87.

2007a. 'Parallel and comparable corpora: the state of play', in Y. Kawaguchi, T. Takagaki, N. Tomimori and Y. Tsuruga (eds.) *Corpus-based Perspectives in Linguistics*, pp. 131–45. Amsterdam: John Benjamins.

2007b. 'Parallel and comparable corpora: what is happening?', in G. Anderman and M. Rogers (eds.) *Incorporating Corpora: Translation and the Linguist*, pp. 18–31. Clevedon: Multilingual Matters.

McEnery, T., Ivanic, R., Smith, N. and Ormerod, F. 1997. 'Multimedia corpora', in B. Lewandowska-Tomasyk and J. Melia (eds.) *Practical Applications of Language Corpora*, pp. 24–33. Łódź University / British Council.

McEnery, T., Wilson, A., Sanchez-Leon, F. and Nieto-Serano, A. 1997. 'Multilingual resources for European languages: contributions of the Crater Project', *Literary and Linguistic Computing* 12 (4): 219–26.

McEnery, T., Tanaka, I. and Botley, S. P. 1997. 'Corpus annotation and reference resolution', in R. Mitkov and B. Boguraev (eds.) *Proceedings of the Association for Computational Linguistics Workshop on Anaphora Resolution for Unrestricted Texts*, pp. 67–74. Madrid. Available online at: www.aclweb.org/anthology/W/W97/W97–1310.pdf.

McEnery, T., Baker, P. and Hardie, A. 2000a. 'Assessing claims about language use with corpus data-swearing and abuse', in J. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, pp. 45–55. Amsterdam: Rodopi. (Reprinted in G. Sampson and D. McCarthy (eds.), 2004. *Corpus Linguistics: Readings in a Widening Discipline*, pp. 396–403. London and New York: Continuum.)

2000b. 'Swearing and abuse in modern British English', in B. Lewandowska-Tomaszczyk and P. Melia (eds.) *PALC'99: Practical Applications in Language Corpora*, pp. 37–48. Berlin: Peter Lang.

McEnery, T., Xiao, R. Z. and Mo, L. 2003. 'Aspect marking in English and Chinese: using the Lancaster Corpus of Mandarin Chinese for contrastive language study', *Literary and Linguistic Computing* 18 (4): 361–78.

McEnery, T., Xiao, R. Z. and Tono, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

McIntyre, D., Bellard-Thomson, C., Heywood, J., McEnery, T., Semino, E. and Short, M. 2004. 'Investigating the presentation of speech, writing and thought in spoken British English: a corpus-based approach', *ICAME Journal* 28: 49–76.

McKoon, G. and Macfarland, T. 2000. 'Externally and internally caused change of state', *Language* 76 (4): 833–58.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum.

Mahlberg, M. 2007a. 'A corpus stylistic perspective on Dickens' *Great Expectations*', in M. Lambrou and P. Stockwell (eds.) *Contemporary Stylistics*, pp. 19–31. London: Continuum.

2007b. 'Clusters, key clusters and local textual functions in Dickens', *Corpora* 2 (1): 1–31.

2007c. 'Corpus stylistics: bridging the gap between linguistic and literary studies', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora: Theory and Analysis*, pp. 219–46. London: Continuum.

Mair, C., Hundt, M., Leech, G. and Smith, N. 2002. 'Short-term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora', *International Journal of Corpus Linguistics* 7 (2): 245–64.

Marcus, M. Santorini, B. and Marcinkiewicz, M. 1993. 'Building a large annotated corpus of English: the Penn treebank', *Computational Linguistics* 19 (2): 313–30.

Mason, O. 1999. 'Parameters of collocation: the word in the centre of gravity', in J. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, pp. 267–80. Amsterdam: Rodopi.

2001. *Programming for Corpus Linguistics*. Edinburgh University Press.

Matthews, D., Lieven, E., Theakston, A. and Tomasello, M. 2004. 'The role of frequency in the acquisition of English word order', *Cognitive Development* 20 (1): 121–36.

Mautner, G. 2009. 'Corpora and critical discourse analysis', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 32–46. London: Continuum.

Meyer, C. 2002. *English Corpus Linguistics: An Introduction*. Cambridge University Press.

Meyers, A. 2009. 'Compatibility between corpus annotation efforts and its effect on computational Linguistics', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 105–24. London: Continuum.

Millar, N. 2009. 'Modal verbs in time: frequency changes 1923–2006', *International Journal of Corpus Linguistics* 14 (2): 191–220.

2011. 'The processing of malformed formulaic sequences', *Applied Linguistics* 32 (2): 129–48.

Milroy, L. and Milroy, J. 1992. 'Social network and social class: towards an integrated sociolinguistic model', *Language in Society* 21 (1): 1–26.

Mindt, D. 1996. 'English corpus linguistics and the foreign language teaching syllabus', in J. Thomas and M. Short (eds.) *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*, pp. 232–47. London: Longman.

Mohamad-Ali, A. 2007. 'Semantic fields of problem in business English: Malaysian and British journalistic business texts', *Corpora* 2 (2): 211–39.

Mohan, P. and Zader, P. 1986. 'Discontinuity in a life cycle', *Language* 62 (2): 291–320.

Moisl, H. and Jones V. 2005. 'Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods', *Literary and Linguistic Computing* 20 (supplement): 125–46.

Mollin, S. 2007. 'The Hansard hazard: gauging the accuracy of British parliamentary transcripts', *Corpora* 2 (2): 187–210.

Monaghan, P. and Christiansen, M.H. 2010. 'Words in puddles of sound: modelling psycholinguistic effects in speech segmentation', *Journal of Child Language* 37 (3): 545–64.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford University Press.

Mougeon, R. and Nadasdi, T. 1998. 'Sociolinguistic discontinuity in minority language communities', *Language* 74 (1): 40–55.

Mukherjee, J. 2004. 'Corpus data in a usage-based cognitive grammar', in K. Aijmer and B. Altenberg (eds.) *The Theory and Use of Corpora: Papers from the Twenty-Third International Conference on English Language Research on Computerized Corpora (ICAME 23)*, pp. 85–100. Amsterdam: Rodopi.

2005. *English Ditransitive Verbs: Aspects of Theory, Description and a Usage-based Model*. Amsterdam: Rodopi.

2006. 'Corpus linguistics and English reference grammars', in A. Renouf and A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*, pp. 337–54. Amsterdam: Rodopi.

2010. 'Corpus linguistics versus corpus dogmatism: *pace* Wolfgang Teubert', *International Journal of Corpus Linguistics* 15 (3): 370–8.

Mukherjee, J. and Gries, St. Th. 2009. 'Collostructional nativisation in New Englishes: verb-construction associations in the International Corpus of English', *English World-Wide* 30 (1): 27–51.

Murphy, B. 2009. '"She's a *fucking* ticket": the pragmatics of FUCK in Irish English – an age and gender perspective', *Corpora* 4 (1): 85–106.

Nelson, G., Wallis, S. and Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Nevalainen, T. and Raumolin-Brunberg, H. 1996. *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

Nevalainen, T., Taavitsainen, I., Pahta, P. and Korhonen, M. 2008. *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*. Amsterdam: John Benjamins.

Nirenburg, S., Somers, H. and Wilks, Y. (eds.) 2003. *Readings in Machine Translation*. Cambridge, MA: MIT Press.

Nokkonen, S. 2006. 'The semantic variation of NEED TO in four recent British English corpora', *International Journal of Corpus Linguistics* 11 (1): 29–71.

Núñez Pertejo, P. 2006. 'An interview with Geoffrey Leech', *Atlantis* 29 (1): 143–56.

Oakes, M. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Oakes, M. and Farrow, M. 2007. 'Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries', *Literary and Linguistic Computing* 22 (1): 85–99.

Oakes, M. and McEnery, T. 2000. 'The background to parallel corpus alignment', in T. McEnery, S. Botley and A. Wilson (eds.) *Multilingual Corpora in Teaching and Research*, pp. 1–37. Amsterdam: Rodopi.

O'Dwyer, B. 2006. *Modern English Structures: Form, Function and Position*. New York: Broadview Press.

O'Halloran, K. and Coffin, C. 2004. 'Checking overinterpretation and underinterpretation: help from corpora in critical linguistics', in C. Coffin, A. Hewings and K. O'Halloran (eds.) *Applying English Grammar: Corpus and Functional Approaches*, pp. 275–97. London: Arnold.

O'Keefe, A. and McCarthy, M. 2010. *The Routledge Handbook of Corpus Linguistics*, London: Routledge.

Orpin, D. 2005. 'Corpus linguistics and critical discourse analysis: examining the ideology of sleaze', *International Journal of Corpus Linguistics* 10 (1): 37–61.

Pang, B. and Lee, L. 2008. 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval* 2 (1/2): 1–135.

Papp, F. 1966. *Mathematical Linguistics in the Soviet Union*. The Hague: Mouton.

Paradis, C. and Lacharité, D. 1997. 'Preservation and minimality in loanword adaptation', *Journal of Linguistics* 33 (2): 379–430.

Partington, A. 2004. '"Utterly content in each other's company": semantic prosody and semantic preference', *International Journal of Corpus Linguistics* 9 (1): 131–56.

Pawley, A. and Syder, F. H. 1983. 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', in J. C. Richards and R. W. Schmidt (eds.) *Language and Communication*, pp. 191–226. London: Longman.

Peacock, M. 2006. 'A cross-disciplinary comparison of boosting in research articles', *Corpora* 1 (1): 61–84.

Petrovic, S., Snajder, J., Basic, B. and Kolar, M. 2006. 'Comparison of collocation extraction measures for document indexing', *Journal of Computing and Information Technology* 14 (4): 321–7.

Phillips, M. 1989. *Lexical Structure of Text*. University of Birmingham.

Piao, S. 2000. 'Sentence and word alignment between Chinese and English', unpublished PhD thesis, Lancaster University.

2002. 'Word alignment in English–Chinese parallel corpora', *Literary and Linguistic Computing* 17 (2): 207–30.

Piattelli-Palmarini, M. (ed.) 1980. *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. London: Routledge and Kegan Paul.

Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P. and Archer, D. 2008. 'The identification of spelling variants in English and German historical texts: manual or automatic?', *Literary and Linguistic Computing* 23 (1): 65–72.

Plag, I., Dalton-Puffer, C. and Baayen, H. 1999. 'Morphological productivity across speech and writing', *English Language and Linguistics* 3 (2): 209–28.

Pollock, E. 2006. *Stalin and the Soviet Science Wars*. Princeton, NJ: Princeton University Press.

Popper, K. R. [1934] 2006. *Logic of Scientific Discovery*. (First English edition published 1959.) New York: Routledge.

Prentice, S. and Hardie, A. 2009. 'Empowerment and disempowerment in the Glencairn uprising: a corpus-based critical analysis of Early Modern English news discourse', *Journal of Historical Pragmatics* 10 (1): 23–55.

Pullum, G. K. and Scholz, B. C. 2002. 'Empirical assessment of stimulus poverty arguments', *The Linguistic Review* 19 (1–2): 9–50.

Quinn, A. and Porter, N. 1996. 'Developing the ICE Corpus Utility Program', in S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English*, pp. 79–91. Oxford: Clarendon Press.

Quirk, R. 1957. 'Relative clauses in educated spoken English', *English Studies* 38 (1): 97–109.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1972. *A Grammar of Contemporary English*. London: Longman.

1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rayner, K. 1998. 'Eye movements in reading and information processing: 20 years of research', *Psychological Bulletin* 124 (3): 372–422.

Rayson, P. 2008. 'From key words to key semantic domains', *International Journal of Corpus Linguistics* 13 (4): 519–49.

Rayson, P., Leech, G. and Hodges, M. 1997. 'Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus', in *International Journal of Corpus Linguistics* 2 (1): 133–52.

Rayson, P., Wilson, A. and Leech, G. 2002. 'Grammatical word class variation within the British National Corpus sampler', in P. Peters, P. Collins and A. Smith (eds.) *New Frontiers of Corpus Research: Papers from the Twenty-First International Conference on English Language Research on Computerized Corpora, Sydney 2000*, pp. 295–306. Amsterdam: Rodopi.

Rayson, P., Archer, D., Piao, S., and McEnery, T. 2004. 'The UCREL semantic analysis system', in *Proceedings of the Workshop on Beyond Named Entity Recognition:*

Semantic labelling for NLP Tasks in Association with the LREC 2004, pp. 7–12. Lisbon, Portugal.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. 2007. 'Tagging the Bard: evaluating the accuracy of a modern POS tagger on Early Modern English corpora', in *Proceedings from Corpus Linguistics 2007*. University of Birmingham. Available online at: www.corpus.bham.ac.uk/conference/proceedings.shtml.

Reed, A. 1978. *CLOC User Manual*. University of Birmingham.

Rissanen, M. 2008. 'Corpus linguistics and historical linguistics', in A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics: An International Handbook*, pp. 53–68. Berlin and New York: Walter de Gruyter.

Rissanen, M., Kytö, M. and Heikkonen, K. (eds.) 1997. *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles*. Berlin: Mouton de Gruyter.

Renouf, A. 2003. 'WebCorp: providing a renewable data source for corpus linguists', in S. Granger and S. Petch-Tyson (eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, pp. 39–58. Amsterdam: Rodopi.

2007. 'Corpus development 25 years on: from super-corpus to cyber-corpus', in R. Facchinetti (ed.) *Corpus Linguistics Twenty-Five Years On: Selected Papers of the Twenty-Fifth International Conference on English Language Research on Computerised Corpora*, pp. 27–50. Amsterdam: Rodopi.

Reppen, R. 2009. 'English language teaching and corpus linguistics: lessons from the American National Corpus', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 204–13. London: Continuum.

Rickford, J. R., Wasow, T. A., Mendoza-Denton, N. and Espinoza, J. 1995. 'Syntactic variation and change in progress: loss of the verbal coda in topic-restricting *as far as* constructions', *Language* 71 (1): 102–31.

Roberts, S. J. 1998. 'The role of diffusion in the genesis of Hawaiian Creole', *Language* 74 (1): 1–39.

Robinson, A. 2009. *Lost Languages: The Enigma of the World's Undeciphered Scripts*. London: Thames and Hudson.

Rock, F. 2001. 'Policy and practice in the anonymization of linguistic data', *International Journal of Corpus Linguistics* 6 (1): 1–26.

Römer, U. 2005. *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: John Benjamins.

Rumelhart, D. E. and McClelland, J. L. 1987. 'Learning the past tenses of English verbs: implicit rules or parallel distributed processing', in B. MacWhinney (ed.) *Mechanisms of Language Acquisition*, pp. 194–248. Mahwah, NJ: Lawrence Erlbaum.

Sagerstad, Y. 2005. 'Changing cultures of written communication: letter – email – sms', in R. Harper, L. Palen and A. Taylor (eds.) *The Inside Text: Social, Cultural and Design Perspectives in SMS*, pp. 9–32. Dordrecht: Springer.

Sampson, G. R. 1987. 'The grammatical database and parsing scheme', in R. Garside, G. Leech and G. Sampson (eds.) *The Computational Analysis of English*, pp. 82–96. Harlow: Longman.

1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

1997. 'Depth in English grammar', *Journal of Linguistics* 33 (1): 131–51.

2000. *CHRISTINE Corpus*, Stage I: Documentation. Available online at: www.grsampson.net/ChrisDoc.html.

2002. 'Regional variation in the English verb qualifier system', *English Language and Linguistics* 6 (1): 17–30.

Sampson, G. R. and Babarczy, A. 2008. 'Definitional and human constraints on structural annotation of English', *Natural Language Engineering* 14 (4): 471–94.

Sampson, G.R. and McCarthy, D. (eds.) 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.

Savoy, J. and Gaussier, E. 2010. 'Information retrieval', in N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing* (second edition), pp. 455–84. Boca Raton, FL: CRC Press.

Sawyer, P., Rayson, P. and Garside, R. 2002. 'REVERE: support for requirements synthesis from documents', *Information Systems Frontiers Journal* 4 (3): 343–53.

Schiffrin, D. 1985. 'Conversational coherence: the role of *well*', *Language* 61 (3): 640–67.

Schmid, H.-J. 2003. 'Do women and men really live in different cultures? Evidence from the BNC', in A. Wilson, P. Rayson and T. McEnery (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, pp. 185–221. Frankfurt am Main: Peter Lang.

Schmitt, N. 2004. *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.

Schütze, H. 1995. 'Distributional part-of-speech tagging', in *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, pp. 141–8. 27–31 March 1995. Dublin, Ireland.

Scott, M. 1996. *WordSmith Tools*. Oxford University Press.

Semino, E. 2008. *Metaphor in Discourse*. Cambridge University Press

Semino, E., Koller, V., Hardie, A. and Rayson, P. 2009. 'A computer-assisted approach to the analysis of metaphor variation across genres', in J. Barnden, M. Lee, J. Littlemore, R. Moon, G. Philip and A. Wallington (eds.) *Corpus-based Approaches to Figurative Language: A Corpus Linguistics 2009 Colloquium*, pp. 145–54. University of Birmingham.

Semino, E. and Short, M. 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Seretan, V. and Wehrli, E. 2007. 'Multilingual collocation extraction with a Syntactic Parser', *Language Resources and Evaluation* 43 (1): 71–85.

Shastri, S. V., Patilkulkarni, C. T. and Shastri, G. S. 1986. *Manual of Information to Accompany the Kolhapur Corpus of Indian English, for Use with Digital Computers*. Bergen: ICAME. Available online at: http://khnt.hit.uib.no/icame/manuals/kolhapur/index.htm.

Short, M., Semino, E. and Culpeper, J. 1996. 'Using a corpus for stylistics research: speech and thought presentation', in J. Thomas and M. Short (eds.) *Using Corpora in Language Research*, pp. 110–31. London: Longman.

Shortall, T. 2007. 'The L2 syllabus: corpus or contrivance?', *Corpora* 2 (2): 157–85.

Siewierska, A. 1993. 'Syntactic weight versus information structure and word order variation in Polish', *Journal of Linguistics* 29 (2): 233–65.

Siewierska, A. and Hollmann, W. B. 2007. 'Ditransitive clauses in English with special reference to Lancashire dialect', in M. Hannay and G. J. Steen (eds.) *Structural–Functional Studies in English Grammar*, pp. 83–102. Amsterdam and Philadelphia: John Benjamins.

Siewierska, A., Xu, J. and Xiao, R. Z. 2010. 'Bang-le yi ge da mang (offered a big helping hand): a corpus study of the splittable compounds in spoken and written Chinese', *Language Sciences* 32 (4): 464–87.

Simpson, R., Briggs, S., Ovens, J. and Swales, J. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor: The Regents of the University of Michigan.

Sinclair, J. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

1991. *Corpus, Concordance, Collocation*. Oxford University Press.

1992. 'The automatic analysis of corpora', in J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, pp. 379–97. Berlin and New York: Mouton de Gruyter.

1996a. *Preliminary Recommendations on Corpus Typology*, Expert Advisory Group on Language Engineering Standards (EAGLES). Available online at: www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpustyp.ps.gz.

1996b. 'The search for units of meaning', *Textus* 9 (1): 75–106.

1999. *Concordance tasks*. Available online at: www.twc.it/happen.html.

2004a. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

2004b. 'Intuition and annotation: the discussion continues', in K. Aijmer and B. Altenberg (eds.) *Advances in Corpus Linguistics*, pp. 39–60. Amsterdam: Rodopi.

Sinclair, J., Jones, S. and Daley, R. 1970. *The OSTI Report*. Available in reprint as Sinclair *et al*. (2004).

Sinclair, J., Jones, S., Daley, R. and Krishnamurthy, R. 2004. *English Collocational Studies: The OSTI Report*. London: Continuum.

Smith, N. and Rayson, P. 2007. 'Recent change and variation in the British English use of the progressive passive', *ICAME Journal* 31: 107–37.

Smith, N., McEnery, T. and Ivanic, R. 1998. 'Issues in transcribing a corpus of children's handwritten projects', *Literary and Linguistic Computing* 13 (4): 312–29.

Snyder, W. and Stromswold, K. 1997. 'The structure and acquisition of English adverb constructions', *Linguistic Inquiry* 28 (2): 281–317.

Somers, H. (ed.) 2003. *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins.

St Clair, M. C., Monaghan, P. and Christiansen, M. H. 2010. 'Learning grammatical categories from distributional cues: flexible frames for language acquisition', *Cognition* 116 (3): 341–60.

Stefanowitsch, A. and Gries, St. Th. 2003. 'Collostructions: investigating the interaction between words and constructions', *International Journal of Corpus Linguistics* 8 (2): 209–43.

2005. 'Covarying collexemes', *Corpus Linguistics and Linguistic Theory* 1 (1): 1–43.

2008. 'Channel and constructional meaning: a collostructional case study', in G. Kristiansen and R. Dirven (eds.) *Cognitive Sociolinguistics*, pp. 129–52. Berlin and New York: Mouton de Gruyter.

Stenström, A.-B., Andersen, G. and Hasund, I. K. 2002. *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings*. Amsterdam: John Benjamins.

Stern, C. and Stern, W. 1907. *Die Kindersprache*. Leipzig: Barth.

Stewart, D. 2009. *Semantic Prosody: A Critical Evaluation*. London: Routledge.

Stuart, K. and Trelis, A. 2006. 'Collocation and knowledge production in an academic discourse community', in C. Pérez-Llantada, R. Alastrué and C.-P. Neumann (eds.)

Proceedings of the Fifth International Conference of the European Association of Languages for Specific Purposes, pp. 238–45. University of Zaragoza.

Stubbs, M. 1993. 'British traditions in text analysis: from Firth to Sinclair', in M. Baker, F. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, pp. 1–36. Amsterdam: John Benjamins.

1994. 'Grammar, text, and ideology: computer-assisted methods in the linguistics of representation', *Applied Linguistics* 15 (2): 201–23.

1995. 'Collocations and semantic profiles: on the cause of the trouble with quantitative studies', *Functions of Language* 2 (1): 23–55.

1996. *Text and Corpus Analysis*: *Computer Assisted Studies of Language and Culture*. Oxford: Blackwell.

1997. Review of T. McEnery and A. Wilson (1996), *Corpus Linguistics. International Journal of Corpus Linguistics* 2 (2): 296–302.

2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

2005. 'Conrad in the computer: examples of quantitative stylistic methods', *Language and Literature* 14 (1): 5–24.

2007. 'On texts, corpora and models of language', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora: Theory and Analysis*, pp. 127–62. London: Continuum.

Sun, C.-F. and Givón, T. 1985. 'On the so-called SOV word order in Mandarin Chinese: a quantified text-study and its implications', *Language* 61 (2): 329–51.

Svartvik, J. (ed.) 1990. *The London–Lund Corpus of Spoken English: Description and Research*. Malabar, FL: Krieger Publishing Company.

Swales, J. M. 2002. 'Integrated and fragmented worlds: EAP materials and corpus linguistics', in J. Flowerdew (ed.) *Academic Discourse*, pp. 153–67. London: Longman.

Tagliamonte, S. 2007. 'Representing real language: consistency, trade-offs and thinking ahead!', in J. Beal, K. Corrigan and H. Moisl (eds.) *Using Unconventional Digital Language Corpora*, Vol. I: *Synchronic Corpora*, pp. 205–40. Basingstoke: Palgrave Macmillan.

Taylor, A., Marcus, M. and Santorini, B. 2003. 'The Penn treebank: an overview', in A. Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*, pp. 5–22. Dordrecht: Kluwer Academic Press.

Taylor, C. 2008. 'What is corpus linguistics? What the data says', *ICAME Journal* 32: 179–200.

Taylor, L. and Barker, F. 2008. 'Using corpora in language testing', in N. H. Hornberger (ed.) *The Encyclopedia of Language and Education*, Vol. VII: *Language Testing and Assessment*, pp. 241–54. New York: Springer.

Temperley, D. 2003. 'Ambiguity avoidance in English relative clauses', *Language* 79 (3): 464–84.

Templin, M. C. 1957. *Certain Language Skills in Children: Their Development and Certain Interrelationships*. Institute of Child Welfare monographs 26. Minneapolis: University of Minnesota Press.

Teubert, W. 2004. 'Language and corpus linguistics', in M. Halliday, W. Teubert, C. Yallop and A. Čermácová (eds.) *Lexicology and Corpus Linguistics: An Introduction*, pp. 73–112. London: Continuum.

2005. 'My version of corpus linguistics', *International Journal of Corpus Linguistics* 10 (1): 1–13.

2007a. '*Parole*-linguistics and the diachronic dimension of the discourse', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, Discourse and Corpora: Theory and Analysis*, pp. 57–88. London: Continuum.

2007b. 'Sinclair, pattern grammar and the question of *hatred*', *International Journal of Corpus Linguistics* 12 (2): 223–48.

2010. 'Our brave new world?', *International Journal of Corpus Linguistics* 15 (3): 354–8.

Teubert, W. and Čermáková, A. 2004. *Corpus Linguistics: A Short Introduction*. London: Continuum.

Theakston, A. L., Lieven, E. V. M., Pine, J. M. and Rowland, C. F. 2001. 'The role of performance limitations in the acquisition of verb-argument structure: an alternative account', *Journal of Child Language* 28 (1): 127–52.

Thelwall, M. 2008. 'Fk yea I swear: cursing and gender in MySpace', *Corpora* 3 (1): 83–107.

Thompson, H. S. and McKelvie, D. 1997. 'Hyperlink semantics for standoff markup of read-only documents', *Proceedings of SGML Europe*. Available online at: www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Tiersna, P. M. 1982. 'Local and general markedness', *Language* 58 (4): 832–49.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Toivanen, I. 2002. 'The directed motion construction in Swedish', *Journal of Linguistics* 38 (2): 313–45.

Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Tono, Y. 2009. 'Integrating learner corpus analysis into a probabilistic model of second language acqusition', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 184–203. London: Continuum.

Tribble, C. 1999. 'Genres, keywords, teaching: towards a pedagogic account of the language of project proposals', in L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Berlin: Peter Lang.

Trudgill, P. 1999. *The Dialects of England*. Oxford: Blackwell.

Unicode Consortium 2006. *The Unicode Standard, Version 5.0*. London: Addison-Wesley.

Ungerer, F. and Schmid, H. J. 1996. *An Introduction to Cognitive Linguistics*. (Second edition published 2006.) London: Longman.

Valera, S. 1998. 'On subject-orientation in English *–ly* adverbs', *English Language and Linguistics* 2 (2): 263–82.

Van Den Heuvel, T. 1988. 'TOSCA: an aid for building syntactic databases', *Literary and Linguistic Computing* 3 (3): 147–51.

van Halteren, H. (ed.) 1999. *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers.

Váradi, T. 2001. 'The linguistic relevance of Corpus Linguistics', in P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL Technical Papers vol. 13, pp. 587–93. Lancaster University: UCREL.

Vepstas, L. 2010. 'Structure in linguistics', *International Journal of Corpus Linguistics* 15 (3): 363–9.

Véronis, J. 2005. Contribution to the *Corpora* list, 12th May 2005. Available online at: http://mailman.uib.no/public/corpora/2005-May/000970.html.

Voutilainen, A. and Järvinen, T. 1995. 'Specifying a shallow grammatical representation for parsing purposes', in *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 210–14. Dublin. San Francisco: Morgan Kaufmann Publishers.

Wasow, T. 2002. *Postverbal Behaviour*. Stanford, CA: CSLI Publications.

Watson, G. J. 1994. 'A multidimensional analysis of style in Mudrooroo Nyoongah's prose works', *Text* 14 (2): 239–85.

1995. 'Multi-dimensional analyses of style in prose literature: a response to Biber', *Text* 15 (3): 371–7.

Weisser, M. 2009. *Essential Programming for Linguistics*. Edinburgh University Press.

Whaley, L. 1997. *An Introduction to Language Typology: The Unity and Diversity of Language*. Thousand Oaks, CA: Sage.

Whitsitt, S. 2005. 'A critique of the concept of semantic prosody', *International Journal of Corpus Linguistics* 10 (3): 283–305.

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.) 1997. *Teaching and Language Corpora*. London: Longman.

Wierzbicka, A. 1972. *Semantic Primitives*. Frankfurt am Main: Athenäum Verlag.

1980. *Lingua Mentalis: The Semantics of Natural Language*. Sydney: Academic Press.

Wilson, A. and Rayson, P. 1993. 'Automatic content analysis of spoken discourse: a report on work in progress', in C. Souter and E. Atwell (eds.) *Corpus-based Computational Linguistics*, pp. 215–26. Amsterdam: Rodopi.

Wittenburg, P. Levinson, S., Kita, S. and Brugman, H. 2002. 'Multimodal annotations in gesture and sign language studies', *Proceedings of LREC 2002*, pp. 176–82. Paris: European Language Resources Association.

Wong, M. 2006. 'Corpora and intuition: a study of Mandarin Chinese adverbial clauses and subjecthood', *Corpora* 2 (2): 187–216.

Woods, A., Fletcher, P. and Hughes, A. 1986. *Statistics in Language Studies*. Cambridge University Press.

Wools, D. 1998. *Multiconcord*. Birmingham: CFL Software Development.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press.

Wulff, S., Stefanowitsch, A. and Gries, St. Th. 2007. 'Brutal Brits and persuasive Americans: variety-specific meaning construction in the *into*-causative', in G. Radden, K.-M. Köpcke, T. Berg and P. Siemund (eds.) *Aspects of Meaning Construction*, pp. 265–81. Amsterdam and Philadelphia: John Benjamins.

Wynne, M. 2005. 'Archiving, distribution and preservation', in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 71–8. Oxford: Oxbow Books.

Xiao, R. Z. 2008. 'Theory-driven corpus research: using corpora to inform aspect theory', in A. Lüdeling and M. Kyto (eds.) *Corpus Linguistics: An International Handbook*, pp. 987–1008. Berlin: Mouton de Gruyter.

Xiao, R. Z. and McEnery, T. 2004a. *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins.

2004b. 'A corpus-based two-level model of situation aspect', *Journal of Linguistics* 40 (2): 325–63.

2005. 'Two approaches to genre analysis: three genres in modern American English', *Journal of English Linguistics* 33 (1): 62–82.

2006. 'Collocation, semantic prosody and near synonymy: a cross-linguistic perspective', *Applied Linguistics* 27 (1): 103–29.

2008. 'Negation in Chinese: a corpus-based study', *Journal of Chinese Linguistics* 36 (2): 274–330.

2010. *Corpus-based Contrastive Studies of English and Chinese*. London: Routledge.

Xiao, R. Z. and Yue, M. 2009. 'Using corpora in translation studies: the state of the art', in P. Baker (ed.) *Contemporary Approaches to Corpus Linguistics*, pp. 237–62. London: Continuum.

Xu, R., Lu, Q. and Li, Y. 2003. 'An automatic Chinese collocation extraction algorithm based on lexical statistics', in *Proceedings of the NLPKE Workshop*, pp. 321–6. Beijing, China.

Yadava, Y. P., Hardie, A., Raj, R., Lohani, R., Regmi, B. N., Gurung, S., Gurung, A., McEnery, T., Allwood, J. and Hall, P. V. 2008. 'Construction and annotation of a corpus of contemporary Nepali', *Corpora* 3 (2): 213–25.

Youmans, G. 1991. 'A new tool for discourse analysis: the vocabulary management profile', *Language* 67 (4): 763–89.

Index