Research and Practice in Applied Linguistics
Series Editors: Christopher N. Candlin and David R. Hall

Corpora and Language Education

Lynne Flowerdew

Research and Practice in Applied Linguistics

General Editors: Christopher N. Candlin and David R. Hall, Linguistics Department, Macquarie University, Australia.

All books in this series are written by leading researchers and teachers in Applied Linguistics, with broad international experience. They are designed for the MA or PhD student in Applied Linguistics, TESOL or similar subject areas and for the language professional keen to extend their research experience.

Titles include:

Dick Allwright and Judith Hanks
THE DEVELOPING LANGUAGE LEARNER
An Introduction to Exploratory Practice

Francesca Bargiela-Chiappini, Catherine Nickerson and Brigitte Planken
BUSINESS DISCOURSE

Alison Ferguson and Elizabeth Armstrong
RESEARCHING COMUNICATION DISORDERS

Lynne Flowerdew
CORPORA AND LANGUAGE EDUCATION

Sandra Beatriz Hale
COMMUNITY INTERPRETING

Geoff Hall
LITERATURE IN LANGUAGE EDUCATION

Richard Kiely and Pauline Rea-Dickins
PROGRAM EVALUATION IN LANGUAGE EDUCATION

Marie-Noëlle Lamy and Regine Hampel
ONLINE COMMUNICATION IN LANGUAGE LEARNING AND TEACHING

Annamaria Pinter
CHILDREN LEARNING SECOND LANGUAGES

Virginia Samuda and Martin Bygate
TASKS IN SECOND LANGUAGE LEARNING

Norbert Schmitt
RESEARCHING VOCABULARY
A Vocabulary Research Manual

Helen Spencer-Oatey and Peter Franklin
INTERCULTURAL INTERACTION
A Multidisciplinary Approach to Intercultural Communication

Cyril J. Weir
LANGUAGE TESTING AND VALIDATION

Tony Wright
CLASSROOM MANAGEMENT IN LANGUAGE EDUCATION

Forthcoming titles:

Anne Burns and Helen da Silva Joyce
LITERACY

Christopher N. Candlin and Stephen H. Moore
EXPLORING DISCOURSE IN CONTEXT AND ACTION

Sandra Gollin and David R. Hall
LANGUAGE FOR SPECIFIC PURPOSES

David Cassels Johnson
LANGUAGE POLICY

Marilyn Martin-Jones
BILINGUALISM

Martha Pennington
PRONUNCIATION

Devon Woods
INSTRUCTIONAL STRATEGIES AND PROCESSES IN LANGUAGE EDUCATION

Chris Candlin and Stephen Moore (*editors*)
EXPLORING DISCOURSE IN CONTEXT AND IN ACTION

Research and Practice in Applied Linguistics
Series Standing Order ISBN 978–1–4039–1184–1 hardcover 978–1–4039–1185–8 paperback
(*outside North America only*)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

Also by Lynne Flowerdew

CORPUS-BASED ANALYSES OF THE PROBLEM–SOLUTION PATTERN

NEW TRENDS IN CORPORA AND LANGUAGE LEARNING
(*co-editor with Ana Frankenberg-Garcia and Guy Aston*)

Corpora and Language Education

Lynne Flowerdew
Hong Kong University of Science and Technology, Hong Kong SAR, China

palgrave
macmillan

Softcover reprint of the hardcover 1st edition 2012 978-1-4039-9892-7

First published 2012 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN 978-1-4039-9893-4 ISBN 978-0-230-35556-9 (eBook)
DOI 10.1057/9780230355569

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

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

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

10 9 8 7 6 5 4 3 2 1
21 20 19 18 17 16 15 14 13 12

Contents

List of Figures

List of Tables

General Editors' Preface

Research and Practice in Applied Linguistics is an international book series from Palgrave Macmillan which brings together leading researchers and teachers in Applied Linguistics to provide readers with the knowledge and tools they need to undertake their own practice-related research. Books in the series are designed for students and researchers in Applied Linguistics, TESOL, Language Education and related subject areas, and for language professionals keen to extend their research experience.

Every book in this innovative series is designed to be user-friendly, with clear illustrations and accessible style. The quotations and definitions of key concepts that punctuate the main text are intended to ensure that many, often competing, voices are heard. Each book presents a concise historical and conceptual overview of its chosen field, identifying many lines of enquiry and findings, but also gaps and disagreements. It provides readers with an overall framework for further examination of how research and practice inform each other, and how practitioners can develop their own problem-based research.

The focus throughout is on exploring the relationship between research and practice in Applied Linguistics. How far can research provide answers to the questions and issues that arise in practice? Can research questions that arise and are examined in very specific circumstances be informed by, and inform, the global body of research and practice? What different kinds of information can be obtained from different research methodologies? How should we make a selection between the options available, and how far are different methods compatible with each other? How can the results of research be turned into practical action?

The books in this series identify some of the key researchable areas in the field and provide workable examples of research projects, backed up by details of appropriate research tools and resources. Case studies and exemplars of research and practice are drawn on throughout the books. References to key institutions, individual research lists, journals and professional organizations provide starting points for gathering information and embarking on research. The books also include annotated lists of key works in the field for further study.

The overall objective of the series is to illustrate the message that in Applied Linguistics there can be no good professional practice that isn't based on good research, and there can be no good research that isn't informed by practice.

Christopher N. Candlin and David R. Hall
Macquarie University, Sydney

Acknowledgements

I am greatly indebted to the editors of this series, Chris Candlin and David Hall, for inviting me to write this book and for their support and encouragement throughout the whole process. I have very much appreciated their insightful feedback and suggestions on various drafts. The final text has benefited a great deal from their editorial input.

This book has taken a few years to reach fruition and is a culmination of 15 years working in the area of corpus linguistics. I have very much enjoyed the various corpus linguistic conferences I have attended over the years: TaLC (Teaching and Language Corpora), IVACS (Inter-Varietal Applied Corpus Studies) and ICAME (International Computer Archive of Modern and Medieval English), to name a few. I would also like to thank the participants at these conferences for their stimulating ideas and discussions. Their names are too numerous to mention here but their work is reflected in this book.

Finally, I would like to thank my family for their support and care during some difficult years.

Part I
Key Concepts and Approaches

1
Definition, Purposes and Applications of Corpora

This chapter will:

- Define what a corpus is
- Outline what corpora can tell us about language and examine the shortcomings of corpus-based linguistics
- Briefly overview the long-established and more recent applications of corpora

1.1 Definition of a corpus

Leading researchers in the field of corpus linguistics (e.g. Sinclair 1991; Stubbs 1996; Biber et al. 1998; Hunston 2002) all view a corpus as a collection of authentic language, either written or spoken, which has been compiled for a particular purpose. Most commonly these purposes are purely linguistic, but can also be of a socio-pragmatic nature. As McCarthy (2001: 63) points out, corpora are 'social artefacts', the investigation of which can uncover the socio-pragmatic behaviour of particular discourse communities (Stubbs 1996; Tognini Bonelli 2001). It is also generally agreed that nowadays when reference is made to a corpus, the corpus data are in machine-readable form which can also be accessed electronically for analysis.

Concept 1.1 Criteria for defining a corpus

A corpus consists of authentic, naturally occurring data;
A corpus is assembled according to explicit design criteria;
A corpus is representative of a particular language or genre;
A corpus is designed for a specific linguistic or socio-pragmatic purpose.

An issue which is also of relevance in defining a corpus is size, an aspect which relates, in one way or another, to the four parameters mentioned above, which are themselves interrelated. Large-scale, general-purpose corpora are generally in the range of 100 million to 500 million words, whereas more specialised, genre-related corpora can be from around 50,000 to 250,000 words.

In general, researchers are unanimous in arguing that in principle the more text, the better. Sinclair (1991, 2004a) emphasises that a corpus should be large and that it 'typically contains many millions of words' such that the more likely it is to give an adequate representation of the language and sufficiently multiple occurrences of the items under investigation. In a similar vein, Sampson (2001: 6) states the need for a 'sizeable sample of real-life usage' to ensure there exists adequate evidence for generating or testing hypotheses about the language.

Quote 1.1 Sinclair on definition of a corpus

> A corpus is a collection of naturally occurring language text, chosen to characterize a state or variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are the clues to the lexical structure of the language.
>
> (Sinclair 1991: 171)

Sinclair stresses that corpora should be viewed as collections of millions of words, doubtless because the main purpose of his research is the compilation of general-purpose dictionaries and grammars, which by their very nature necessitate the examination of millions of words to arrive at as complete a description of the system as possible. However, there has been somewhat of a backlash against the notion of size for size's sake, with McCarthy (1998: 23) cautioning thus: 'As long as we keep a cool head in the face of the exhilaration of computer power and vast arrays of text, we will not fall into the temptation of substituting cold numbers for the real people who actually produced the words', thereby reiterating the concept of a corpus as a 'social artefact'.

While as a general rule bigger is probably best, this whole question of what is considered to be an appropriate size for a corpus is highly dependent on the phenomenon one is investigating and the purpose of the corpus enquiry. With regard to the investigation of specific items, McEnery and Wilson (2001) point out that the lower the frequency of the feature one wishes to investigate, the larger the corpus should be. This would apply to nouns, adjectives, adverbs, etc. (i.e. content words) which tend to have a much lower frequency than grammatical words in any given corpus. Conversely, one can argue that smaller corpora can be used for investigating the more common features of

language, such as grammatical items, and indeed Biber (1990) has pointed out that smaller corpora are perfectly adequate for purposes such as these.

Another factor influencing the size of the corpus relates to the degree of internal variation in the language or genre under study. The greater the variation, the more samples and a larger corpus are required to ensure representativeness and thus validity of the data (Meyer 2002). However, here again, size has to be balanced against the level of delicacy of the investigation, an issue touched upon in Kennedy (1998), who remarks on the danger of having too much output such that the data are unwieldy to work with.

A prime example of research dealing with finely grained analyses is that carried out by Carter and McCarthy (1995, 1997) and McCarthy (1998) using the 5-million-word CANCODE (Cambridge and Nottingham Corpus of Discourse in English) of informal registers in English (see Section 3.4). Smallish samples of a few thousand words can yield useful insights into the linguistic realisation of strategic competence for maintaining interpersonal relations. There is thus a case to be made for using more qualitative data for examining very specific sub-purposes concerning socio-pragmatic behaviour, which could easily be overlooked in larger-scale quantitative analyses.

All the above issues concerning the size of a corpus would, in part at least, be determined by the design criteria (cf. Atkins et al. 1992), which are themselves dependent on the purposes of the investigation. Granger (1998a) illustrates how principled design criteria based on extra-linguistic features of text can be applied in learner corpus design, making a distinction between features which relate to the language situation and those characterizing the learner.

Concept 1.2 Learner corpus design criteria

Language	**Learner**
Medium	Age
Genre	Sex
Topic	Mother tongue
Technicality	Region
Task setting	Other foreign languages
	Level
	Learning context
	Practical experience

(Granger 1998a: 8)

Such design criteria are not only applicable in language learning but can also be extended to other fields such as teacher education and translation. For example, in teacher education the language aspect of the corpus design criteria would include considerations as whether the language to be examined was geared towards classroom management or textbook materials. Whether the

practitioners were in-service (novice; experienced) or trainee teachers would be other factors to consider (see Section 7.6).

As Sinclair (1991: 9) notes, 'the results are only as good as the corpus' and it is therefore a *sine qua non* that the above conditions for defining a corpus be met to ensure reliable and valid data for analysis and interpretation by the corpus linguist.

Concept 1.3 Sinclair's ten key principles for developing linguistic corpora

1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.
3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.
4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
6. Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.
7. The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
8. The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
9. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.
10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

(Extracted from Sinclair 2005: 1–21)

1.1.1 Corpus vs database

A corpus is different from, and not to be confused with, a database, or text archive, as several corpus linguists have pointed out (Leech 1991; Hunston 2002; Meyer 2002).

Concept 1.4 Corpus vs database

A corpus is a collection of naturally occurring language, which has been systematically planned and collected in accordance with principled external design criteria with an a priori purpose in mind, which, in turn, determines the design parameters. A database, or text archive, on the other hand, is a large repository of text which is unstructured and often compiled according to what is easily obtainable rather than based on systematic sampling techniques. There is also a difference in the 'reading' of a corpus vs a database: a corpus is read non-linearly whereas it is usually a whole text which is accessed in a database.

However, as noted by Meyer (2002), there remains some dissension as to what exactly constitutes a corpus. Whereas the Expert Advisory Group on Language Engineering Standards (EAGLES) ('Corpus Encoding Standard': http://www.cs.vassar.edu) defines a corpus in a very general way, stating that in addition to newspapers, poetry, drama, etc. it can also contain word lists and dictionaries, Meyer points out that most corpus linguists would probably not go along with such a broad definition. Items such as wordlists of various kinds would constitute databases, according to those working within the traditions of corpus linguistics, as opposed to language processing.

1.1.2 Corpus vs Web

Whether the Web constitutes a corpus has been hotly contested, most notably by Sinclair.

Quote 1.2 Sinclair on the World Wide Web

The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present, it is quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled.

(Sinclair 2005: 4)

However, in spite of Sinclair's misgivings, Kilgariff and Grefenstette (2003) put forward the argument that the Web can be considered a corpus on the following grounds. They maintain that some corpus linguists frame this issue as 'What is a corpus?' when what they should be asking instead is 'Is corpus *x* good for task *y*?' Following from this, they deduce that the Web is a corpus, if *A corpus is a collection of texts when considered as an object of language or literary study* (p. 334). They address the issue of absence of representativeness in Web data, the main criticism levelled against the web, by raising points to show that this criterion is not clearly articulated in the compilation of large-scale general corpora either.

Quote 1.3 What does representativeness mean?

> If we wish to develop a corpus of general English, we may think it should be representative of general English, so we then need to define the population of 'general English-language events' of which the corpus will be a sample. Consider the following issues:
>
> - *Production and reception*: Is a language event an event of speaking or writing, or one of reading and hearing? Standard conversations have, for each utterance, one speaker and one hearer. A *Times* newspaper article has (roughly) one writer and several hundred thousand readers.
> - *Speech and text*: Do speech events and written events have the same status? It seems likely that there are orders of magnitude more speech events than writing events, yet most corpus research to date has tended to focus on the more tractable task of gathering and working with text.
>
> (Kilgariff and Grefenstette 2003: 340–1)

Various advantages of using the Web as a corpus have been put forward. Mair (2006), while acknowledging the 'big-and-messy' approach of using 'dirty' Web data, exemplifies the value of the Web for the study of infrequent and recent linguistic phenomena such as neologisms, and the study of 'World Englishes' (see Section 5.3). Search engines, which interface with commercial search engines, have been specially devised for Web searches to allow users to submit more sophisticated queries which are then presented in formats similar to those of corpus output, e.g. collocation tables. Such search engines include WebCorp (Morley 2006; Renouf et al. 2007), KWiKFinder (Fletcher 2002) and WebPhraseCount (Schmied 2006b).

1.2 Why corpora?

Investigations of corpora, in common with all linguistics, are concerned with the description and explanation of the structure and use of language. However,

two key defining features of corpus-based linguistics are that the analyses are based on empirical data and tend to be associated with the phraseological approach to language.

Concept 1.5 Phraseology

Stubbs (2001a: 59) neatly encapsulates the general concept of phraseology thus: 'the pervasive occurrence of phrase-like units of idiomatic language use'. He further elaborates by pointing out that phraseological units '… have typical components, but are highly variable, with probabilistic relations between the components; they are typically realised by a sequence of several word forms, but their boundaries do not correspond systematically to syntactic units'.

As a general illustration of this phenomenon, Gledhill (2000) shows that the phraseology of *in* in the results sections of cancer research articles is typically indicated by the phrase *increase in* occurring between two noun phrases, with the first one indicating an experimental treatment and the following one a measurable outcome, e.g. *treatment with butyrate resulted in an increase in relative tumor rates*.

A bedazzling plethora of terms abound in the literature for this concept. While some linguists prefer the term phraseology (e.g. Cowie 1998), others opt for formulaic language or sequences (Wray 2002). The term 'multi-word unit' is also commonly found as an alternative. See Wray (ibid.: 9) who lists over 50 different terms to describe this phenomenon.

More explicitly, for Sinclair (2004a) this phraseological approach consists of five categories of co-selection with the core lexical item and semantic prosody as obligatory elements, and collocation, colligation and semantic preference as optional categories. The starting point for analysis is usually with frequency counts of lexical items.

1.2.1 Frequency data

Computationally, it is easy to generate various types of frequency lists. What is at issue here, however, is the decisions that need to be made at the pre-processing stage as to what exactly would qualify as an individual item to be counted in a frequency list regardless of how the list is displayed.

The most basic type of frequency list shows the number of types (individual occurrences of any word form) and their tokens (i.e. frequency) in the corpus. These lists of types and tokens are usually sorted either alphabetically or by rank frequency (summaries of different types of frequency lists can be found in Barnbrook (1996), Bowker and Pearson (2002) and Sinclair (1991)).

Quote 1.4 What does a frequency list show?

> A frequency list for a corpus shows you the words that occur in it and the relative proportion that each contributes towards it. If your corpus is properly representative, this information can give you a reasonably accurate picture of the language as a whole. Problems are likely to arise from the separation in the lists of word forms belonging to the same lemma and from the existence of homographs. Software capable of lemmatizing the lists has been developed, but homographs are normally disambiguated manually.
>
> (Barnbrook 1996: 134–5)

Basic word frequency lists are very useful in the preliminary stages of analysis as they provide a good starting point for more detailed investigation into the grammatical or content words in a text. The word frequency lists in Table 1.1 show the most frequent 25 types (lemmatised; see following section) in a single economics text of about 320,000 words, compared with the most frequent 25 types in a general academic corpus of a similar size. A quick comparison of the two lists reveals that there are seven content words (marked with an asterisk*) in the economics text, whereas there are none in the general academic texts of the top 25 most frequent words. Such frequency data give a bird's-eye view of the composition of the corpora and can suggest useful avenues for further exploration.

To reiterate Sinclair's point that the results are only as good as the corpus, by the same token, a wordlist is only as good as the corpus from which it is derived. In this respect, Sinclair underscores the importance of the size of the corpus for generating frequency data on account of the enormous imbalance in the frequency of words, given that even in a very long text about half of the vocabulary will have occurred only once in the text. He notes that if one wishes to study all the uses of a word (some of which may be highly infrequent) then a corpus needs to contain many million words in order to generate a sufficient number of occurrences of the different uses on which to base descriptive statements.

However, here again, considerations of size have to be tempered by considerations of whether the corpus under investigation is of a general or specialised nature. Sinclair's point above is doubtless true for general corpora, but the frequency lists above suggest that even smallish corpora of around 300,000 words can throw up useful frequency data, provided that they concentrate on a specific subject area or genre. And indeed, Sinclair (2005) provides statistical evidence in support of this point by a comparison of frequency data across two 1-million-word corpora: LOB, a general corpus, and the HKUST (Hong Kong University of Science and Technology) Corpus of English in Computing Science (cf. James et al. 1994).

Example 1.1

Table 1.1 Word frequency lists (lemmatised)

Rank	Economics text		General academic texts	
	Type	**Token**	**Type**	**Token**
1	the	22,905	the	23,890
2	of	12,710	of	14,591
3	be	10,686	be	14,021
4	a	9,952	and	8,455
5	and	8,323	in	8,077
6	in	7,010	to	7,867
7	to	6,502	a	7,857
8	that	4,392	that	3,400
9	price*	3,080	have	3,217
10	for	2,912	this	3,143
11	it	2,674	it	3,017
12	we	2,534	for	3,060
13	have	2,514	as	2,574
14	cost*	2,251	by	2,351
15	by	2,034	with	2,110
16	this	2,003	they	2,056
17	demand*	1,944	on	1,956
18	on	1,882	which	1,631
19	as	1,831	he	1,590
20	they	1,820	not	1,544
21	curve*	1,804	or	1,542
22	at	1,797	at	1,535
23	firm*	1,743	from	1,518
24	supply*	1,590	can	1,242
25	quantity*	1,467	We	1,205

(Adapted from Kennedy 1998: 101)

Lemmatisation

Lemmatisation is a useful procedure for obtaining general information on how common a particular word is in a corpus. Through lemmatisation, different word forms of the same word class are grouped together; these could either be the singular and plural of nouns, e.g. *book, books*, or different forms of the same verb, e.g. *play, plays, played, playing* (although the software would have to be more sophisticated to cope with irregular forms of nouns and verbs).

However, some corpus linguists, most notably Sinclair, have challenged the very concept of a lemma and drawn attention to the potential drawbacks of lemmatisation, which are discussed below.

Concept 1.6 Lemmatisation

In its strictest sense, lemmatisation involves grouping word forms from the same word class under the base or uninflected form of the word. For example, the word forms *take*, *takes*, *took*, *taken*, *taking* and *to take*, from the word class of verb, would all be counted as belonging to the lemma *take*. Sometimes, however, word forms from different word classes, e.g. *surprise* and *surprisingly*, are categorised under the same lemma. As Sinclair (1991) notes, lemmatisation is not a straightforward process as it sometimes involves a degree of subjective judgement on the part of the researcher.

Three challenges have been raised against lemmatisation of individual words. First, if meaning was constant across different inflected forms of a word, then it would be a fairly straightforward matter to lemmatise a corpus. However, this is certainly not the case. Sinclair and Renouf (1988) provide corpus evidence to demonstrate that the morphological pair *certain* and *certainly* behave quite independently of each other in terms of meaning and also usage patterns. On the other hand, Stubbs (1996: 172), taking the lemma *class* by way of example, raises the issue of whether other word forms such as *classify* and *classification*, which are semantically related to some meanings of *class*, should be considered as separate lemmas. At a greater degree of specificity regarding meaning, Tognini Bonelli (2001) questions whether *facing* and *faced* should be assigned to the lemma *face* as the former has a concrete meaning (e.g. *facing forwards*), whereas the latter retains only the metaphorical meaning (e.g. *faced with a dilemma*).

Secondly, even in cases where the meaning does remain constant across different word forms, lemmatisation would of course conceal in a frequency list which forms of a particular lemma are used or are the most common in the corpus. In the wordlist in Table 1.2 (Bowker and Pearson 2002) the wild card* has been used to retrieve all word forms beginning with 'repr-'. Such unlemmatised lists may provide a more helpful indicator for further exploration of the corpus, as to why for example the singular noun *representation* occurs twice as frequently as its plural form (a check of these two word forms in the 100-million BNC revealed the frequencies to be even more marked, with *representation* occurring nearly three times as frequently as its plural form).

Example 1.2

Table 1.2 Wordlist (unlemmatised)

Type	Token
represent	18
representation	8
representations	4
representative	1
representatives	6
represented	3
representing	5
represents	11
repressed	1
reprieve	1
reprocessing	3
reproduce	9
reproduced	2
reproduces	2
reproducing	2
reproduction	6
reproductive	9
reprogram	1
reprogrammed	1
reprogramming	2

(Adapted from Bowker and Pearson 2002: 116)

Third, Sinclair (1991) also casts doubt on the concept of a lemma as being the base form of a word and puts forward the suggestion that the most frequently encountered word form could equally well be regarded as the lemma. This point is well worth considering for corpus application. As an example, in the whole of the BNC the base form *legislate* occurs 199 times, whereas there are 6960 instances of the noun *legislation*. However, this does not necessarily mean that frequency should take precedence in determining what is to be taught (see Section 7.1).

In sum, even the notion of what constitutes a lemma is debatable from a meaning potential point of view and in terms of frequency of occurrence, and hence (in view of these factors) lemmatisation of a corpus may not always be desirable.

Identification of types

Types, usually considered as occurrences of individual words, are very often referred to in relation to tokens, invoking a concern for type/token ratio.

Concept 1.7 Type/token ratio

The type/token ratio is a shorthand expression for the following formula:

Ratio = number of occurrences of type/the total number of tokens in the corpus.

The ratio obtained is usually scaled up (to avoid working with very small numbers) by multiplying the ratio obtained by 100 and expressing the result as a percentage. In such a way the results can be normalised to enable comparison of different corpora.

The type/token ratio is useful for showing the composition of the corpus as a whole. For instance, in Table 1.3 in Example 1.3 of the 10 most frequent words in the *BNC Sampler* (a 2-million-word corpus of speech and writing consisting of 184 samples taken from the BNC), the cumulative type/token ratio indicates that nearly 25 per cent of the corpus as a whole consists of tokens of only 10 different types.

Example 1.3

Table 1.3 Table showing type/token ratio

N	**Type**	**Token**	**Type/token ratio (%)**	**Cumulative %**
1	THE	109,830	5.15	5.15
2	AND	54,759	2.57	7.72
3	OF	50,752	2.38	10.10
4	TO	49,538	2.32	12.42
5	A	42,436	1.99	14.41
6	I	38,974	1.83	16.23
7	IN	35,292	1.65	17.89
8	IT	34,308	1.61	19.50
9	YOU	30,700	1.44	20.94
10	THAT	30,208	1.42	22.35

(Adapted from Scott 2001: 56)

However, this statistical procedure is not without its problems for the same reason as lemmatisation. Just as the definition of a lemma is not clear-cut, the related issue of what constitutes a 'type' is also open to question and needs to be addressed in the pre-processing stage before frequency counts are computed.

One problem arising with the type/token ratio is that decisions need to be made regarding the boundaries of what counts as a 'type' in the corpus. For

example, the researcher needs to decide whether to count abbreviations where separated by a full stop, and apostrophes where used to signal a shortened form of the verb, as one unit, i.e. a single type. This problem is even more accentuated with regard to the phraseological approach to language where one word does not equate with a semantic unit, which may be spread across several words in a fixed or semi-fixed phrase. In cases where the phrase is completely fixed, e.g. *of course*, there is no difficulty, but in others which may allow some internal variation, as in the case of *in fact* which can be expanded to *in actual fact*, the categorisation of types is a somewhat murky area and begs the question as to how far along the continuum of fixedness and semi-fixedness one can or should go in the categorisation of types.

Such semi-fixed phrases are very often assigned a pragmatic function which can lead to further complications when the corpus is analysed from a more discourse-based perspective. A few researchers (e.g. Aijmer 1984) have examined a restricted set of individual hedges such as *sort of, kind of, in a way, somehow*. However, the lexico-grammar realising hedging is extremely subtle, encompassing many different types of lexical verbs and modals, and it is difficult to conceive how the different combinations and permutations expounded in Hyland's (1998) work on hedging could be classified according to types. Would the compound hedge *would appear* and *appear*, marginally differing in their degree of attenuation, be classified as one type and how would one search a corpus for these?

Another problem relates to the status of a word and whether it has a technical or more of a general meaning. Many words of a technical nature, i.e. specific to a particular sub-discipline, often occur as part of a multi-word unit (see Yang 1986 for a corpus investigation of such items). When the same word occurs on its own it could take on a more general meaning. For example, a glance at the wordlist of the economics text in Example 1.1 reveals *price* and *demand* occurring in the top 20 most frequent words. However, it could well be that such words occur as part of a multi-word unit (e.g. *price/earnings ratio, price fixing, supply and demand*) in which case there would be some justification for categorising the combinatorial lexis as one type. Also, as this list is lemmatised it may conceal other specialist multi-word terms such as *pricing policy*. And then we return to the issue of variation, in this case lexical. If *pricing policy* is counted as one type, would *pricing strategy* also be counted as belonging to the same type since *policy* and *strategy* are listed as synonyms in the *Collins English Thesaurus* (1998)?

The problem of what constitutes a type is further compounded when different language systems are considered. Different problems would manifest themselves in agglutinative languages (such as Turkish), languages which combine single words into compound words (such as German) and languages that make heavy use of particles (such as Thai and Japanese).

A question mark therefore hangs over the identification of types when issues of phraseology, pragmatic functions, vocabulary type and language systems are taken into account.

Frequency data vs salience

Both Widdowson (1991, 2003) and Cook (1998) have raised the issue of frequency vs salience (although Widdowson opts for the term *prominence*), an aspect already touched on in the discussion of representativeness (see Quote 1.3).

Quote 1.5 Cook on 'salience'

> Even as a record of 'facts' computer corpora are incomplete. They contain information about production but not about reception. They say nothing about how many people have read or heard a text or utterance, or how may times. Thus a memo hastily skimmed by one person and consigned to the wastepaper basket counts equally with a tabloid headline read by millions. Or with a text, such as a prayer or poem, which is not only often repeated but also deeply valued. Occurrence, distribution, and importance, in other words, are not the same. This applies to short texts, but also to shorter units. Some phrases pass unnoticed precisely because of their frequency, others strike and stay in the mind, though they may occur only once. And because different individuals notice different things, such saliency can never be included in a corpus.
>
> (Cook 1998: 59)

Salience is discussed from two different perspectives in the literature: cognitive and cultural. Discussing the former in terms of prototypicality, Widdowson (2003: 83–4) illustrates how 'prototypical prominence in the mind does not accord with frequency of actual occurrence'. Citing a study by Rosch in which respondents were asked to give a hyponym for the superordinate category 'vegetable', Widdowson notes that the word 'cauliflower' received a higher prototype score than the word 'potato', but this difference in cognitive representation did not correspond with measures of textual frequency.

Approaching this mentalist type of prototypical prominence from a grammatical starting point, Shortall (2005) elicited data on the *there*-construction from native and non-native speakers in English, and from Taiwanese speakers in Mandarin. His findings were that in over 70 per cent of subject sentences, the prototype there + be + NP + PP was produced (e.g. *There is a book on the table*). This provided a stark contrast to the figure of 32 per cent for this pattern in the 450-million-word Bank of English, with the corpus data showing a wide distribution over six different *there*-construction patterns. Also, *there*-constructions were found to occur more frequently with abstract nouns (e.g. *There is some evidence*

to suggest that …) whereas the respondents favoured concrete nouns. Although Shortall tentatively concluded in his presentation that 'prototypes seem to reflect a language instinct that works similarly across different languages', it may be that the answer is to be found in the background of the learners. It could well be that the students, especially low-level learners, produced this form because of exposure to it most likely through textbooks or because of other unknown factors such as the influence of interlanguage, grammatical complexity, etc.

Why an item is culturally salient is easier to account for than cognitive salience as usually it is reasonably obvious why a particular phrase or item resonates with the listener or reader. Hunston (2001: 195) cites examples of slogans, proverbs, headlines, advertisements, and Cook (1998) poems and prayers, as all having cultural salience. Devices associated with creative exploitation of language such as play on words, alliteration, repetition, mimicry, etc. all serve to make a striking impression in one way or another.

Example 1.4 Cultural salience

Too many musical heroes	spoil the broth,	but not on Bill Laswell's late
cordon bleu chef might just	spoil the broth.	I don't think anybody really
workers: Too many computers	spoil the broth.	WASHINGTON, DC
Will one more TV cook	spoil the broth?	Not if it's TODAY columnist
penicillin, too many cooks	spoil the book	When every other
part of PR. Too many cooks	spoil the menu;	There's a recipe for

(Hunston 2001: 195)

Interestingly, Hunston cites several examples of cultural salience where salience and frequency counts do appear to coincide, a convergence not noted between cognitive salience and frequency data.

One of the most striking examples of a culturally salient phrase was the advertisement by Saatchi & Saatchi for Margaret Thatcher's 1978 prime ministerial campaign in the UK. The headline 'Labour Isn't Working' was accompanied by a print advertisement showing a long, winding line outside an unemployment office. This advertisement achieved resonance through the clever thought-provoking interplay of words and illustration. However, as Cook (1998: 5) rightly points out, 'Even as a record of "facts" computer corpora are incomplete.' In such cases involving advertising media, corpora are indeed incomplete as they can only provide restricted information on printed output, and give us no idea as to how many times the advert was aired and where and in how many print sources it appeared. Also, as Cook notes, '[corpora] say nothing about how many people have read or heard a text or utterance, or how many times'. (This issue can be related to the challenges of defining

the concept of 'representativeness' in general corpora; see Quote 1.4.) Cook's observation on reception can be extended to considerations of illocutionary force (i.e. if the reader understands the text in the way it was intended) and perlocutionary effect (i.e. the action which the text provoked the reader to take). Presumably this advert was effective on both counts, as it was, in fact, credited as a major factor in Thatcher's electoral win. These are the kinds of concerns which occupy critical discourse analysts; see for example, Fairclough (2000) who critically examines the political rhetoric of New Labour (i.e. the discursive strategies employed by the Labour government in Britain to promote its ideology through 'media spin'), to uncover the real meaning underlying the manipulative strategies employed to influence public perception. Corpus data thus fall short on the side of production in certain domains, and also especially with regard to the conditions of reception, both of which have implications for the concept of representativeness discussed earlier.

1.2.2 Collocational data

Statistical vs textual collocation

A corpus is regarded as indispensable for observing collocational patterning, visible through the vertical display of the node word (i.e. the word form or lemma under investigation, commonly known as KWIK, keyword in context) and its collocate(s), i.e. word or words with which it co-occurs. It is generally agreed that collocational behaviour of a lexical nature can be manifested up to a span of four or five words to the left or right of the node word. However, one area of disagreement among linguists concerns how collocation is to be recognised. Although Cowie and Howarth (1996) make a case for viewing collocation as a 'textual' phenomenon on the grounds that certain restricted collocations may only show up infrequently in a corpus and may be subject to arbitrary variation, the majority of researchers (Sinclair 1991; Stubbs 1996; Hunston 2001) favour the identification of collocations through various statistical means.

Quote 1.6 Cowie and Howarth on collocation

> Collocations are often described as fixed and recurrent word-combinations. But both parts of this description are misleading. Typically, collocations are not fixed but variable to a limited and arbitrary degree. As for frequency, it can be shown that individual restricted collocations may recur to only a limited extent within a given text or across several texts devoted to the same topic. It is best to think of a collocation as a familiar (institutionalized), stored (memorized) word-combination with limited and arbitrary variation.
>
> (Cowie and Howarth 1996: 82)

Although Cowie and Howarth (1996) seem to view collocations as acting as multi-word units with 'limited and arbitrary variation', the display of multiple concordance lines of the node word when 'read vertically' shows that collocational choices are not purely arbitrary, operating according to the slot-and-filler model of a paradigmatic, substitution-like table, but in fact are highly patterned, probabilistic choices which form a linear syntagmatic relation. Further statistical treatment of the data can also show that collocations operate syntagmatically, as opposed to paradigmatically, by providing numerical evidence for the strength and certainty of a collocation. For example, Hunston (2001), using the 450-million-word Bank of English, shows that the strongest (i.e. the top) adverbial collocates for *significant* are *radiologically, statistically, electorally, militarily* and *symbolically*, with *statistically* also being a certain collocate (i.e. the probability of its co-occurrence is high based on the frequency of this word in the corpus as a whole).

The notion that co-occurrence of words is a highly patterned syntagmatic relation is further reinforced by reference to work examining collocational patterning in relation to a particular lemma. Stubbs (1996) demonstrates that even different forms of the same lemma are restricted in their collocational behaviour; as regards the lemma *educate*, the form *education* mainly collocates with terms denoting institutions (e.g. *further, higher, university*) whereas the form *educate* is found to collocate with the near synonyms *enlighten*, *help* and *inform*. In fact, Sinclair (1992) has proposed that examining the collocational patterns of different word forms can be used as a basis for deciding whether they belong to the same lemma or not; similar behaviour would mean that the word forms qualify for categorisation under the same lemma. By extension, the fact that different forms of the same lemma can exhibit quite different collocational behaviour is an argument against lemmatisation of different word forms of the same base form (see discussion on lemmatisation in Section 1.2.1).

Another important aspect of collocation is that the items may be separated by other non-fixed or semi-fixed words and may have a different position relative to one another. Methodologies have been devised which can account for this type of variability (see Cheng et al. 2008a and Greaves 2009 for an account of the ConcGram software).

Example 1.5 Variability in collocations

Durrant (2009: 158) describes a statistical method using *WordSmith Tools* (Scott 1999) which can identify both constituency and positional variation of the collocation in the sentence *He made a* ***powerful argument***, e.g.:

> He made a *powerful*, but ultimately unconvincing, *argument*.
> His *argument* was a *powerful* one.

Collocational variability in engineering texts, which has been extracted using the ConcGram software, is described in Warren (2010).

However, as Cowie and Howarth point out, statistical measurements cannot always be used for determining collocation, especially in cases where word combinations 'may recur only to a limited extent'. By way of illustration, Stubbs (2001a: 74–5) cites the example of a small corpus yielding the following data for the node adverb 'distinctly':

> <distinctly <N + 1: cagey, cool, dated, dour, downbeat, iffy, inferior, meaner, muted, strange, thin, unimpressed, unwell>

In the above case, as the adjectival collocates occurred only once each, statistical measures to determine the likelihood of co-occurrence could not be carried out. Although the corpus does not yield statistically significant results regarding purely *lexical* collocation, it does yield patterning at the *semantic* level, with 'distinctly' shown to have an attraction for disapproving words. It is in instances such as these found in small-scale corpora where a 'textual' approach to collocation rather than a purely 'statistical' one would seem to be more helpful (see Section 3.1.2 for discussion of a 'psycholinguistic' approach to collocation).

Semantic prosody

The final part of the previous section briefly touched on the concept of semantic prosody, first dealt with in depth in a seminal and oft quoted article by Louw (1993), who acknowledges Sinclair (1987, 1991) as generating the first computationally derived 'profile' of this phenomenon (Sinclair, himself, refers to Bréal (1897) for coining the term 'contagion' to signify the transference of meaning, usually pejorative, as the product of habitual collocations).

Concept 1.8 Hunston on semantic prosody

The features of semantic prosody can be summarised thus:

- The semantic prosody of a lexical item is a consequence of the more general observation that meaning can be said to belong to whole phrases rather than to single words.
- Semantic prosody can be observed only by looking at a large number of instances of a word or phrase, because it relies on the *typical* use of a word or phrase.
- It accounts for 'connotation': the sense that a word carries a meaning in addition to its 'real' meaning. The connotation is usually one

> of evaluation, that is, the semantic prosody is usually negative or, less frequently, positive.
> - It can be exploited, in that a speaker can use a word in an atypical way to convey an ironic or otherwise hidden meaning.
> - The semantic prosody of a word is often not accessible from a speaker's conscious knowledge. Few people, for example, would define *SET in* as meaning 'something bad starts to happen', but when the negative connotation is pointed out in many cases it accords with intuition (*A spell of fine weather set in* sounds very odd, for example).
>
> (Hunston 2002: 142)

While semantic prosody seems to be a relatively straightforward concept to pin down according to its main features summarised above, various research studies suggest that the 'connotation' a word carries in addition to its 'real' meaning may not always be easy to assign. Although Stubbs (1995a, 2001b) presents clear-cut evidence, based on an analysis of 40,000 examples across 120 million words of the Cobuild Corpus, to show that CAUSE is typically associated with nouns indicating 'something bad' (*anxiety, cancer*, *problems*) whereas PROVIDE collocates with mostly 'good things' (*care, food, help, money*), Kjellmer's (2005) data demonstrate that verbs do not always fall so neatly into a positive or negative polarity category, and concludes that semantic prosody could be seen as more of a probabilistic phenomenon rather than a plus-or-minus one. Using the same Cobuild Corpus, Kjellmer examined the pattern verb + bare noun, where the noun is the direct object (e.g. *abandon ship*). The examples presented show that in this pattern while verbs such as *add* and *allow* have a positive alignment in 47 and 35 per cent of cases respectively, they also display neutrality in 43 and 65 per cent of cases. Also, positive polarity was more in evidence than negative polarity in Kjellmer's data, which is somewhat surprising given that other corpus linguists (Louw 1993; Hunston 2001) have found that semantic prosody is usually negative.

Prosodic colouring can also vary, depending on which corpora are consulted; see Tribble (2000) who proposes that the semantic prosodies of a word may be both 'universal' and 'local', i.e. a word may be seen to have a global prosody in relation to the language as a whole, but a localized prosody in a specific discipline or genre. For example, Partington (2004a) illustrates how LAVISH is very often used with nouns to indicate disapproval in newspaper reporting, whereas its use in normal British conversation does not have this prosody. These examples bring us to one of Whitsitt's (2005) critiques of the concept of semantic prosody. Whether a word can be said to be *imbued* with a particular semantic prosody is open to question, as illustrated below.

Quote 1.7 Whitsitt's critique of semantic prosody

> … there is no evidence for assuming that we can see the results of a diachronic process of *imbuing*. … The contrary, in fact, seems to be the case. One need but consider verbs like *alleviate, heal, relieve, soothe* etc., all perfect candidates for semantic prosody since they habitually appear in the company of clearly unpleasant words, yet it seems clear that a word like *alleviate*, to take one example, certainly does not come to have an unpleasant meaning because of that company.
>
> (Whitsitt 2005: 297)

As illustrated so far, semantic prosody is usually considered in terms of lexical relations, but there are a few reports in the literature where colligation is involved (see Section 1.2.3). For example, Sinclair (1998) notes that the use of 'the' + adjective which is used to refer to the whole class of people described by the adjective, is commonly followed by evaluative adjectives of a negative nature such as 'the elderly', 'the unemployed', 'the sick', etc., a point which grammars fail to point out. Moreover, Louw (1993) has illustrated that the semantic prosody of certain verbs can change, depending on whether they are used transitively or intransitively. For example, where *build up* is used transitively, with a subject denoting people, the prosody is uniformly good, with *build up* followed by objects such as *organisations, understanding*, etc. However, where used intransitively with subjects such as *cholesterol, toxins* and *armaments*, the prosody is invariably bad.

Another issue raised in cases where a word is used literally or metaphorically concerns whether the prosody associated with the literal meaning carries over into the metaphorical use of the word and vice versa. Hunston (2002: 120), using examples of the metaphorical expressions *turn a blind eye to* and *turn a deaf ear to*, whose meanings are to be construed as a conscious avoidance strategy, states that there is no evidence to suggest that the prosodic meaning associated with *blind* and *deaf* in these phrases carries over into the literal meanings of blindness and deafness. In contrast, Sinclair's (1999) example of *budge* exemplifies how the prosody of the literal meaning, indicating frustration or irritation on the part of the person who wants someone or something moved, is very similar to the prosody of the metaphorical meaning. In the second example below, Sinclair, citing Louw (personal communication), shows with reference to the wider co-text how *budge* is being used both literally and metaphorically; co-textual words including the use of *office* and *sitting* signal that moving the President from his office simultaneously requires both his physical ousting and cancellation of his presidential authority:

> he recognizes it, he'll refuse to budge off that stool where he's sitting n
> ight me out of his mind and refuse to budge. In that case, the Vice-President

The typical semantic prosody of an item may therefore not always be so easy to determine when the type of corpus (general vs specialised), colligational features and metaphorical/literal meaning are taken into account.

However, one of the most contentious issues is whether semantic prosody is a semantic or pragmatic phenomenon, which has implications for corpus interpretation. Sinclair (2004a: 34) points out that semantic prosody carries attitudinal meaning (as do Hunston and Stubbs), and is 'on the pragmatic side of the semantics/pragmatics continuum'. Widdowson (2004), on the other hand, seems to view prosody more as a semantic phenomenon recoverable from co-occurrences of collocations in the text. He agrees with the basic notion of semantic prosody, in that being a co-textual relation, semantic signification can be read off from concordance lines, but also views the assignation of pragmatic significance as somewhat problematic on the grounds that it does not exist in the text, but can only be ascertained from other co-textual and contextual features.

Quote 1.8 Widdowson on semantic signification vs pragmatic significance

> … on the evidence of their customary collocates, particular words can be shown to have a typical positive or negative semantic prosody, and it can be plausibly suggested that facts of co-textual co-occurrence should be recognized as part of the semantic signification of such words. But this, of course, does not tell us about what pragmatic significance might be assigned to such a co-occurrence in a particular text. The point about these co-textual findings is that they are a function of analysis, with texts necessarily reduced to concordance lines. One might trace a particular line back to its text of origin, but then if it is to be interpreted, it has to be related not to other lines in the display but to the other features of the original text.
>
> (Widdowson 2004: 60)

However, in a paper on corpus semantics Stubbs (2001b) argues that the conventionalised view that pragmatic meanings are usually inferred by the reader/listener may be overstated and that large-scale corpus studies can provide evidence to show that pragmatic meanings can also be conventionally encoded in linguistic form. (In fact, Stubbs prefers the term 'discourse prosody' to 'semantic prosody' on account of the fact that prosodic information can often only be established by looking beyond the concordance line to more discourse-based extended units of meaning.) Evidence in support of Stubbs' point is provided by the research of O'Halloran and Coffin (2004) who show how an *accumulation* [my italics] of negative co-texts for *United States of Europe* in a 45-million-word corpus of the British newspaper *The Sun* display a regular

negative attitude for 'United States of Europe'. When encountered as individual instances these examples may appear fairly neutral to someone unfamiliar with the anti-Euro stance of *The Sun* newspaper and its vocabulary.

Example 1.6 Negative semantic prosody

towards their ambition of a	United States of Europe,	stretching from Shetland
could pave the wave for a	United States of Europe.	British people have made
leader's bleak plan for a	United States of Europe	came as a hammer blow to
the road towards a Federal	United States of Europe.	Hague has never tried to
forming into a giant	United States of Europe	– with the same tax and
for a hopeless dream of a	United States of Europe.	He is certain to pay the

(Adapted from O'Halloran and Coffin 2004: 288)

Where one is dealing with multiple texts from a particular discourse community with somewhat institutionalised practices or known entrenched ideologies, it may be possible, as Stubbs has noted, to establish pragmatic meanings based on regularly co-occurring instances of conventionalised attitudes in the text. Whether pragmatic significance can be read off from concordance lines may therefore be dependent on what contextual background knowledge analysts bring to interpretation of the corpus data. However, as Widdowson (2004: 60) points out, such prosodic information does not tell us about what pragmatic significance might be assigned to a co-occurrence of items in just one particular text, and for such information it would be necessary to interpret the item with reference to not only other co-textual features, but also to external contextual information.

Semantic preference

Quote 1.9 Stubbs on semantic preference

> Semantic preference is the relation, not between individual words, but between a lemma or word-form and a set of semantically related words, and often it is not difficult to find a semantic label for the set…. Another example is the word-form *large*, which often co-occurs with words for 'quantities and sizes'.
>
> (Stubbs 2001a: 65)

Semantic preference, sometimes also referred to as semantic association, is closely related to the concept of semantic prosody. The relationship between these two phenomena has often been described as one of set (preference) and subset (prosody). However, the difference is not always so distinct and it

may be more helpful to consider them as interacting in a kind of symbiotic relationship constrained by, and constraining each other, with semantic prosody operating at a more discourse-based level than semantic preference as the following example illustrates (see Partington 2004a for further examples).

Concept 1.9 Relationship between semantic prosody and semantic preference

Stubbs (2001a: 89–95) discusses the behaviour of UNDERGO, which favours various semantic preferences. To the right, it collocates with items from the semantic sets of medicine (e.g. *treatment, surgery*), tests (e.g. *examination, training*) and change (e.g. *dramatic changes*). To the left, it collocates with words expressing some type of involuntariness (e.g. *must, forced to, required to*). At the same time, these preferences, on both the left and right of UNDERGO, determine and are determined by the semantic prosody, in this case negative, i.e. people are forced involuntarily to undergo changes or training, which are very often unfavourable (e.g. *dramatic changes, rigorous test*).

The same points that have been raised for semantic prosody can also be applied to semantic preference. First, preference for a particular semantic domain can be inextricably bound to the grammatical environment in which it occurs. As Partington notes, when CAUSE is followed by a single object, this is frequently an illness, e.g. *Smoking causes cancer*. However when two objects are involved the second object is usually some kind of unpleasant feeling or emotion, e.g. *causes them inconvenience*.

Secondly, the semantic preference of a lexical item may well vary depending on the type of corpus which is searched. For instance, Nelson (2006) compared the semantic associations of *global* in a 1-million-word business English corpus, BEC, with its semantic set of nouns in the BNC. It was found that while *global* was very rich in terms of semantic sets in BEC, e.g. *global products*, *global economic indicators*, in the BNC it was limited to two semantic sets relating to climate, e.g. *global warming*, and people, e.g. *global consumer, global viewer*.

In common with Kjellmer (2005), Partington (2004a) views semantic prosody as more of a probabilistic phenomenon, which he discusses in relation to semantic preference. He provides corpus evidence to demonstrate that items belonging to the same semantic set may have different degrees of negative prosody: in the HAPPEN set, SET IN has the worst prosody, followed by HAPPEN, then OCCUR and TAKE PLACE, with COME ABOUT seeming to be neutral. Moreover, these shadings may differ across languages (see Xiao and McEnery (2006) for a cross-linguistic analysis of semantic prosody, drawing on data from English and Chinese).

The general issues raised in this section regarding semantic prosody and semantic preference have important implications for professional discourses, as outlined below.

Concept 1.10 Implications of semantic prosody and semantic preference for professional discourses

Practitioners working in the fields of media studies and political speech writing might appreciate how semantic prosody can be manipulated on the part of the speech writer, and how it can be misconstrued by an audience. Partington (2003: 231), using a 250,000-word corpus of around 50 press briefings from the late Clinton years, shows how a spokesperson exploits this device for political gain. In the example below, the unusual collocation of *perpetrate* with *truth*, a verb which usually collocates with pernicious activities (e.g. *fraud, bullying*), succeeds in conveying the message that the spokesperson is referring to political enemies, without spelling it out.

> (61) MR LOCKHART: I think they accurately reflect that there are people in Belgrade, in Yugoslavia who understand what the truth here is, as opposed to those who seek to *perpetrate something other than the truth*.

In translation work, it may not be sufficient to know whether certain words have a general tendency towards positive or negative semantic prosody, but also to know *under what circumstances*, i.e. grammatical environment, discourse type, literal vs metaphorical meaning, etc. It is also necessary to consider *to what extent* an item carries an inherently positive or negative semantic prosody, as there are cases where items belonging to the same semantic set have different degrees of negative prosody. Moreover, literal translations may evoke a negative prosody in one language but not in another. Sinclair (2004a: 34) cites the example of the Italian, *a occhio nudo*, a literal translation of *the naked eye*, which has a prosody correlated with some kind of difficulty in English, but not in Italian.

The notion of 'faux amis' type words ('actuellement' in French does not translate as 'actually' in English, but means 'now') can also be extended to 'false semantic prosodies'. For example, as Partington (1998: 77–8) notes, the adjective *impressive* in English has a positive semantic prosody, collocating with words such as *achievement, talent, gains*, etc, whereas in Italian *impressionante* was found to collocate with neutral or unfavourable happenings, e.g. *price rises, assassination attempts*. As Partington (1998: 78) cautions, 'The pitfalls for translators unaware of such prosodic differences are evident.'

More recently, semantic prosody and preference have been considered from the standpoint of Hoey's (2005) theory of lexical priming, which is further explicated in the following section on colligational data.

Concept 1.11 Hoey's theory of lexical priming

Hoey's theory holds that an individual attaches particular meanings to words and phrases, not just based on their intrinsic meanings or connotations, but also based on the previous contexts in which they have habitually encountered them. Priming operates at an individual level, but members of the same discourse communities will share similar primings.

For example, Hoey (2005: 25–6) notes the following for the semantic associations for *consequence* in his corpus made up of 95 million words of *Guardian* news and feature text, supplemented by 3 million words from the BNC (written text) and 230,000 words of spoken data. Four main categories of semantic association were found to constitute 90 per cent of all adjectives premodifying *consequence*. The largest of the semantic associations of *consequence* (59 per cent) was a class of adjective alluding to the underlying process described:

> Whatever his decision, it will be seen as a **logical** *consequence* of a steady decline in influence.

The second largest semantic association (15 per cent) constituted adjectives negatively evaluating the consequence:

> The **doleful** *consequence* is that modern British society has been intensely politicised.

The third semantic association (11 per cent) was made up of adjectives expressing a view regarding the seriousness of the consequence:

> The most **serious** *consequence* of this crime has been the effect on the children.

The fourth semantic association (6 per cent) related to adjectives expressing the 'unexpectedness' of the consequence:

> Yet that very process brought its own **surprising** *consequence*.

Hoey comments thus: '… if the corpus reflects an individual's experience of reading the *Guardian* (or perhaps other) newspaper, then the word *consequence* will be primed for the reader in such a way that they will expect it to occur with such associations' (p. 26).

1.2.3 Colligational data

Corpus-based studies of various aspects of collocation discussed above have somewhat eclipsed work on colligation, a term first coined by Firth (1957: 13) to denote 'the grammatical company a word keeps' (cited in Hoey 1997: 8), and most commonly viewed as a type of relation bridging lexis and grammar. Like collocation, colligation is usually viewed as a probabilistic phenomenon rather than a term for describing rule-governed language, e.g. which verbs are followed by an infinitive or 'that' clause. Hoey's definition of colligation given below and corpus findings will act as the pivot for the discussion in this section.

Quote 1.10 Hoey on colligation

A definition of colligation

The basic idea of colligation is that just as a lexical item may be primed to co-occur with another item, so also it may be primed to occur in or with a particular grammatical function. Alternatively it may be primed to avoid appearance in or co-occurrence with a particular grammatical function….

For current purposes, I suggest that *colligation* can be defined as:

1. the grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank;
2. the grammatical functions preferred or avoided by the group in which the word or word sequence participates;
3. the place in a sequence that a word or word sequence prefers (or avoids).

(Hoey 2005: 43)

Hoey's first definition of colligation, i.e. the grammatical company that a word keeps, can be demonstrated by the finding that *reason*, with the sense of *cause*, shows a strong preference for demonstrative deictics, with a tendency to avoid possessive deictics (Hoey 1997). This avoidance strategy also seems to operate in the case of synonyms or near-synonyms. Comparing the colligational profiles of *actual* and *real*, Tognini Bonelli (1993) notes that *actual* is preceded by the definite article in 99 per cent of cases, whereas only 15 per cent of the instances of *real* are found in this syntactic patterning.

Hoey's second definition of colligation relates to whether the word occurs in the subject, object, complement, etc. slot. By way of example, corpus evidence presented in Francis (1991) demonstrates that certain nouns (e.g. *accident, context, independence*) have their own grammar as they were found to be

unevenly distributed across subject, object and complement slots, as illustrated by *accident* below. This concept harks back to Harris' (1954) work on defining categories in terms of their distributional evidence.

Concept 1.12 The 'grammar' of *accident*

	Subject	Object	Complement	Indirect object/ object complement	Adjunct	Qualifier
accident	8	9	31	—	49	3

The syntagmatic environment, i.e. colligational preference, of *accident* is often an adjunct, e.g. … *wife died as a result of an industrial accident or disease*, with 22 of its 49 adjunct occurrences accounted for by the set phrase 'by accident'. Rarely does it occur as subject or object, and when it does, as subject, its collocational preference is with 'occur' or 'happen' and when as object it is usually associated with 'have' or 'cause'. *Accident* occurs 31 times as complement, often in modalised expressions, discussed in terms of whether the event was or was not one, e.g. *It wasn't an accident at all*; … *thinks the collapse may not have been an accident*.

(Adapted from Francis 1991: 146–7)

The grammatical functioning of a word or phrase can also be considered from a Hallidayan perspective and whether it occurs in the Theme or Rheme position. Of interest is that different senses of a word have been found to have different colligational Theme/Rheme patterning. For example, Hoey (1996) shows that *consequence* with the meaning of 'importance' (always found in his data in the pattern *of … consequence* in Rheme position) is primed to avoid Thematised position so as to avoid potential confusion with *consequence* in the sense of 'result', which occurs frequently in Theme, as illustrated below:

the CMA was an evangelical organisation *of consequence*.
The *consequence* inevitably will be that more commuters will abandon …

All the corpus examples cited above exist at the interface of lexis and grammar at the *sentential* level, but work by Hoey (2004a, b, 2005) considers colligation from a more *textlinguistic* perspective, by exemplifying how certain phrases are primed (i.e. associated with certain positions in a sentence or text built up from previous encounters with that word). Hoey (2004b: 38) shows how *sixty* collocates with *years* which also collocates with *ago* and that this phrase has a strong colligation with Theme, a positive colligation with paragraph initial position when thematised, and a positive colligation with text initial position, when paragraph initial.

1.2.4 Lexico-grammatical patterning

Stubbs' (1996: 36) well-known phrase 'There is no boundary between lexis and grammar: lexis and grammar are interdependent' aptly conveys this convergence between the two. It is self-evident from the above discussions that in the lexico-grammatical approach meaning and usage, uncovered through recurring concordance examples, have primacy, i.e. language is viewed as a 'grammaticalised lexis' rather than a 'lexicalised grammar'. It can also readily be seen that the starting point for linguistic analysis tends to be with the lexical item. One issue arising from this observation, though, is that in some cases it may be more appropriate to start with the grammatical category. (This point is further explored in Section 3.1.1.)

However as Stubbs (2001a: 242) points out, '… it will take a long time before an appropriately revised division of labour between an expanded lexis and reduced syntax has been worked out'. Physicists have been trying for decades to reconcile through string theory models two seemingly incompatible areas in their discipline: quantum mechanics (governing atoms and all other particles) and general relativity (interaction of matter and gravity). The groundbreaking research of corpus linguists such as Sinclair, Stubbs, Hoey and Hunston has set the road map for a more linear, syntagmatic approach to language analysis, where vocabulary and grammar are seen not as two polarities but brought together as a unified whole. Corpus linguistic methodologies have indeed been invaluable in showing how inextricably bound together vocabulary and grammar are, but there is still a way to go in the working out of an all-encompassing theory to offset the main limitations outlined below.

1.3 Why not corpora?

1.3.1 Summary of main limitations

In the previous section I focused on the numerous insights into the patterning of language that are to be gained by employing corpus-based methodologies. Criticisms and limitations of corpus research can be found in Hunston (2002: 22–3) and Partington (1998: 144–9), with rebuttals to some common objections reported in Stubbs (2001a: 221–6) and L. Flowerdew (2005). A synthesis of the main objections together with their rebuttals is presented below.

Concept 1.13 Limitation of size

- **Corpora can never be fully representative as they are delimited by size**

For Chomsky (cited in Stubbs 2001a) a corpus is finite and always 'skewed'; for this reason, it can never be fully representative of the language as a whole (see Section 2.2 for other criticisms levelled against corpus linguistics by the Chomskyan camp). However, Biber (1990) argues that for many of the more

common features of language relatively modest general corpora do provide adequate evidence. Furthermore, results can be checked across corpora for verification of the findings to minimise any inherent bias (Stubbs 2001a).

Specialised corpora (e.g. research articles on biochemistry), in contrast to general corpora, are sampled from a homogeneous population, which could be seen as ensuring representativeness. By virtue of their homogeneous nature, specialised corpora contain a far fewer number of different word forms (types) than general corpora (Sinclair 2005), and for these reasons size is not such an issue.

Concept 1.14 Limitation of contextual features

- **A corpus presents decontextualised language data divorced from its original context.** (Aston 1995; Widdowson 1998)

This statement is almost always certainly true as far as large-scale general corpora are concerned. However, in the case of small, specialised corpora where the analyst is more often than not also the compiler, the analyst often has recourse to the original context, and the corpus can be marked up for these different contextual features, e.g. setting, text type, communicative purpose, etc. (cf. L. Flowerdew 2004).

Concept 1.15 Limitation of software tools

- **Corpus analyses are limited by tools which take a 'bottom up', i.e. a micro-view of the text, rather than a 'top down' approach** (Swales 2002, 2004), **and that due to these technological restrictions, corpus investigations over-emphasise single word forms and collocations at the expense of more discourse-based features of language.**

Concordance software does allow the user to scan up and down a stretch of text and to examine the wider co-text, in terms of sentences and paragraphs, which is often necessary. Such manual searching also allows the analyst to discern 'inter-collocations', a network of words and phrases whose prosodies support each other and contribute to textual cohesion (Stubbs 2001a). Stubbs illustrates how inter-collocations of the verbs ACCOST and LURK examined in the Cobuild corpus build up a network showing common ways of talking about threats to public safety, especially in the mass media (p. 205):

- Accosted – in the street – by a stranger – lurk beneath the surface – hidden beneath the surface – danger – warning

Although, as Stubbs points out, there exists no tool for automatic identification of such intertextual lexical fields, the *Keyword* function in *WordSmith*

Tools (Scott 1997, 1999), which identifies associates (words of very high frequency) occurring in the same text as the item under investigation, is a very promising way forward for establishing the kind of intertextual relations Stubbs has in mind. Moreover, advances in software, such as *WMatrix* (Rayson 2008), allows for searching of semantic fields.

1.3.2 Difficulties with interpretation

One of the most problematic aspects concerning corpus data is that of *interpretation*, which is also compounded by the limitations outlined above. As Hunston (2002) cautions, interpretations of corpus data are deductive in nature and two thorny issues arise regarding extrapolations: possibility and acceptability of a word or grammar structure searched in a corpus, and the sociopragmatic decoding of text.

Possibility and acceptability

Section 1.2 emphasised the probabilistic nature of corpus data. Although frequency counts may reveal typical patterns, they leave a lot unsaid by what is omitted (cf. Section 1.2.1 which has shown that they are not a reliable indicator of 'salience'). They present only positive data, and cannot tell us whether something is *possible* or not. A word or expression may be possible, but not recorded in the corpus. Positive data can confirm intuitions about language, but the lack of data cannot be used as evidence for refuting an intuition.

Moreover, even when positive data are recorded, a corpus might throw up language which may not be acceptable to prescriptive grammarians being a non-standard form of English, may be incorrect structurally, or as is often the case with the English of advanced learners, sound a bit odd to native speakers (Pawley and Syder 1983). For example, in the BNC the following example for *recommend* is found, which, according to the grammar warning note in the *Longman Dictionary of Contemporary English* (p. 1371), the construction *recommend you* + *to*-infinitive, would be classed as a mistake.

> I am doing the firm no harm by tipping you off like this because Kovacs recommends you to buy a clock from them …

However, it is highly likely that practitioners sympathetic to the concept of English as a lingua franca would accept this non-standard form (see Section 6.1). Another murky area where decoding of corpus data can be fraught with problems is in relation to the pragmatic decoding of text.

Socio-pragmatic decoding of text

Widdowson (1998, 2002, 2004) has commented in various places on the lack of contextualisation accompanying corpus data: 'The text travels, but the context does not travel with it', as mentioned in Concept 1.14. While corpus data can be very useful for highlighting specific functions, their pragmatic import is less clear. The corpus data below from a 1-million-word business letters corpus (http://ysomeya.hp.infoseek.co.jp) show that the collocation *cordial invitation* is most commonly used when the sender is offering the invitation, e.g. *Please accept our cordial invitation to visit …*, whereas *kind invitation* is used for thanking, e.g. *Thank you for your kind invitation to attend …*

Example 1.7 Functional vs pragmatic decoding of collocational patterning

Thanking you for your	kind invitation	to address the audience, I rem
Thank you very much for your	kind invitation	to your Christmas party.
Many thanks for your	kind invitation	to join in your celebrations on
I thank you for your	kind invitation	on the occasion of your openin.

May we extend to you a	cordial invitation	to call in at White's and make a
I would like to extend my	cordial invitation	to you to visit our Tokyo office
Province of <name> extends a	cordial invitation	to <name> attend a reception on
This is a	cordial invitation	to become one of Myer's charge

(Adapted from L. Flowerdew 2009: 399)

Such corpus data may only be able to provide some help with interpretation at the pragmalinguistic level (the collocation of certain linguistic features in a certain register). A student may be able to glean from the co-text of *kind invitation* the knowledge that 'I thank you for your kind invitation on the occasion of …' belongs to a more formal register than 'Many thanks for your kind invitation to join in …'. However, the intended illocutionary force, which is influenced by social, cultural and personal preferences and the dynamics of the unfolding interaction, may be more difficult to discern. Here, the use or non-use of certain direct or indirect speech acts can pose problems, especially in cross-cultural situations (Thomas 1983). It could well be that the example 'Province of <name> extends a cordial invitation to <name> to attend …' is a polite directive disguised as an invitation. But we have no way of knowing this unless we are familiar with the situation and social roles of the participants in that particular context (see discussion on semantic prosody in Section 1.2.2 for which the same kind of difficulty with interpretation may arise).

In sum, although corpus data are very useful for revealing typical lexico-grammatical patterns and functional aspects of language, they do have limitations and cannot always be relied upon to tell us what is possible, acceptable, well-formed or pragmatically appropriate.

1.4 General applications of corpus findings

1.4.1 Well-established applications

It was in the late 1980s and early 1990s that large-scale corpora were first used as the basis for creating general dictionaries and grammars, viz. the Collins Cobuild set of reference materials (Sinclair 1987, 1990). A decade later saw the publication of the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). This empirical-based grammar also highlights lexico-grammatical patterns, but its overall descriptive framework is quite different from that of the Cobuild grammar (see Section 7.2.1 for further discussion of these grammars).

Around the same time as the publication of the first edition of the Cobuild dictionary and grammar, pioneering work by Tim Johns (1988, 1991) introduced concordancing techniques into language teaching with a focus on its application in English for academic purposes (see Section 7.2.2 for further pedagogic developments). Interestingly, Johns' 1988 article pre-empts several issues raised more than a decade later: the *credibility* of corpus data vs intuition (see Section 2.2.2) and the *transferability* of the results (see Section 7.1). In the early 1990s learner corpora flourished with the establishment of the International Corpus of Learner English project under the direction of Sylviane Granger (1993) for the purposes of compiling a corpus of argumentative essays from learners of various L2 backgrounds, although the application of learner corpora has been less pronounced.

1.4.2 More recent applications

The imbalance between written and spoken corpora of an academic nature has been somewhat redressed by the compilation of two major spoken corpora, the Michigan Corpus of Spoken Academic English (MICASE) (http://www.lsa.umich.edu/eli/micase/) and its English counterpart, the British Academic Corpus of Spoken English (BASE) (http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/), the findings of which have important pedagogic applications especially in the area of lecture comprehension (see Section 4.4).

Another major development is the broadening out of the area from its concentration on applications of general academic English to take account of various features of English for Specialised Academic Purposes. The compilation of corpora from which test items can be derived is also becoming increasingly popular (see Section 6.8), and the role of corpora in the training of teachers

(see Section 7.6) and the training of translators (see Section 7.5) is gaining momentum. However, what is somewhat surprising is that there has not been more growth in the compilation of corpora of other languages for foreign language teaching, such as those endeavours reported in Section 7.4.1.

Hoey (1998) quoted in Sampson and McCarthy (2004) has remarked: 'Corpus linguistics is not a branch of linguistics, but the route into linguistics.' As this book illustrates, corpus linguistics has without doubt become more interdisciplinary, having made inroads into textlinguistics (see Section 4.3), forensic linguistics (see Section 6.3) and literary stylistics (see Section 6.4).

Quote 1.11 Sampson and McCarthy on the growth of corpus linguistics

> Fifty years ago, corpus linguistics was an obscure and highly specialized minority activity. Since then, slowly at first but in the last ten years almost explosively, it has widened out to provide virtually every approach to the study of language, humanistic or technological, with new methods and new insights. By now, many agree with a widely quoted remark by Michael Hoey in 1998: 'Corpus linguistics is not a branch of linguistics, but the route into linguistics.' This is a good time to become a corpus linguist.
>
> (Sampson and McCarthy 2004: 5)

The above quotation from Sampson and McCarthy reaffirms the status that corpus linguistics has achieved. Although it may now be a 'good time to become a corpus linguist', whose time has now come, the following chapter demonstrates that this has not always been the case.

Further reading

Hunston, S. (2009) Corpus compilation. Collection strategies and design decisions. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics: an International Handbook*, vol. 2 (pp. 154–68). Berlin: Mouton de Gruyter. This article gives a comprehensive overview of the major issues to be considered in corpus design.

Lindquist, H. (2009) *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press. This volume provides a reader-friendly introduction to the main concepts discussed in this chapter.

O'Keeffe, A. and McCarthy, M. (eds) (2010) *The Routledge Handbook of Corpus Linguistics*. Section II: Building and designing a corpus: what are the key considerations? (pp. 31–103). London: Routledge. For readers new to corpus linguistics, this handbook section presents a very user-friendly, uncomplicated introduction to key considerations for building different types of corpora (spoken, written, audio-visual).

Stewart, D. (2009) *Semantic Prosody. A Critical Evaluation*. London: Routledge. This accessible volume draws together all the major work in the area of semantic prosody.

2
Historical and Conceptual Background of Corpus Linguistics: Evolutionary, Revolutionary or Counter-Revolutionary?

This chapter will:

- Present a historical perspective of corpus linguistics
- Show the stages of evolution in corpus linguistics
- Dissect the Chomskyan vs corpus linguistics debate

2.1 Evolutionary

Corpus linguistics can be seen as evolutionary in the sense of developing from Harris' (1954) concept of defining categories in terms of distributional evidence, and Firth's (1957) approach to investigating the meaning of a word by examining its collocations through computational means. To rephrase the title of a seminal article by N. Francis (1992) 'Language Corpora BC', i.e. before the use of computers, the evolution of corpus linguistics can be seen as either BC or AD, i.e. after the advent of digitalisation.

2.1.1 Language corpora BC

Suffice it to say that surveys of pre-electronic corpora can be found in N. Francis (1992), Kennedy (1998: 13–19) and McEnery and Wilson (2001: 2–4) for various fields of scholarship. A few very early key figures could be singled out such as Samuel Johnson for his assembly of over 150,000 citations for the headword entries in his *Dictionary of the English Language* and Otto Jespersen for his 300,000–400,000 slips of paper of striking forms and sentence constructions for the four volumes of his *Modern English Grammar*. More recently, West's (1953) General Service List, a set of 2000 headwords, each representing a word family, has had a wide influence on foreign language teaching and served as the basis for graded readers. It has also inspired a multitude of corpus-based studies on general and disciplinary-specific academic vocabulary (see Section 7.1).

In a somewhat tongue-in-cheek article, McKenny (2003) makes a spirited case for Jonathan Swift (1667–1745) as a precursor for this new way of doing linguistics. Swift's satirical writings were published not long after the establishment of the Royal Society, whose motto 'Nullius in verba' (loosely translated as 'take nobody's word for granted') reflected their principles of acquiring and testing knowledge through empirical investigation. Swift's collection of prefabs and Samuel Johnson's assembly of quotations to illustrate usage in his dictionary, could well be seen as a parallel to developments in scientific enquiry presaging the technological revolution.

Quote 2.1 McKenny on Swift as a precursor of corpus linguistics

Swift had already drawn attention, through the professor of Lagado Academy, to the importance of prefabs in building or reconstituting text. He takes up prefabs again in *Polite Conversation*, this playful work which was the result of a lifetime of cataloguing examples of linguistic abuse. As such, it can therefore be viewed as a continuation of the project outlined in *Proposal to Correct the English Tongue*. He was collecting clichés in order to extirpate them. He describes how he built up a collection of fashionable sayings over 12 years of field work:

> I determined to spend 5 mornings, to dine 4 times, pass three afternoons, and six evenings every week in the houses of the most polite families … I always kept a large table-book in my pocket and as soon as I left the company I immediately entered the choicest expressions.

He then spent a further 16 years 'digesting it into a method'.

(McKenny 2003: 7)

In the 1950s, American structural linguists such as Harris and Fries, for whom a corpus was a *sine qua non* for linguistic description, dominated the scene. For example, Fries (1952) manually compiled a 250,000-word corpus of recorded telephone conversations, the equivalent of the size of small, specialised corpora today, for *The Structure of English*. Then, corpus linguistics, like Rip Van Winkle, went to sleep for 20 years (Leech 1992), and the evolution partly stopped in its tracks with the rise of Chomskyan linguistics of the rationalist variety as a backlash against the prevailing behaviourist theories of the period.

However, in spite of the domination of Chomskyan linguistics in the 1950s and 1960s, corpus linguistics was not completely derailed and the field evolved sporadically with the advent of machine-readable corpora.

2.1.2 Language corpora AD

McEnery (2000) has remarked that the earliest work using machine-readable corpora, predating work carried out at Brown University, was that by Roberto Busa and Alphonse Juilland. Roberto Busa can be considered as the first true corpus linguist, using a concordancer to analyse a corpus of medieval philosophy texts in a collaborative project with IBM running from 1949 to 1967. In fact, it was IBM who first helped create an automatic concordancer for indexing the works of St Thomas Aquinas. Alphonse Juilland, working in the self-described area of 'mechanolinguistics', carried out contrastive corpus linguistics based on frequency lists and produced a frequency dictionary of Spanish words among many other publications. He also addressed issues of balance, covering a wide range of genres, representativeness, examining a range of writers within various genres of 500,000-word corpora and dispersion, developing Harris' concept of distributionalism.

Quote 2.2 McEnery and Wilson on the importance of Busa and Juilland

> Juilland was a pioneer of corpus linguistics. It was Busa and Juilland together who worked out much of the foundations of modern corpus linguistics. If their contributions are less well known, it is largely because corpus linguistics became closely associated with work in English only, and neither Busa nor Juilland worked on corpora of modern English.
>
> (McEnery and Wilson 2001: 20–1)

It was the *Survey of English Usage* (SEU), a 500,000-word corpus of 200 samples of written and spoken texts compiled by Quirk and Greenbaum at University College London, that instigated further work on corpora. In fact, the SEU can be considered as straddling language corpora BC and AD; originally it was not in machine-readable format, but the spoken component was later computerised by Quirk and Svartvik. However, not until the compilation of a 1-million-word corpus at Brown University under the direction of Kučera and Francis (1967) in the 1960s did the era of machine-readable corpora really come into being. The *Brown Corpus* and its British counterpart, the *LOB* (*Lancaster–Oslo/Bergen*) *Corpus*, are generally regarded as the major first-generation corpora. Second-generation mega-corpora include *The Bank of English* (BoE), which currently stands at around 500 million words, and the smaller 100-million-word *British National Corpus* (BNC), for which an American version, the ANC (see Reppen and Ide 2004), is well underway (see Kennedy 1998 for a comprehensive overview of these first- and second-generation corpora).

This then brings us to the question of what constitutes or will constitute third-generation mega-corpora. Undoubtedly, this is the Web, which is assuming increasing prominence and becoming easier to mine and navigate with advances in search engines and tools. However, the Web is not without its pitfalls, and 'provocative questions' about the nature of language on the Web have been raised in Section 1.1.2.

Quote 2.3 Kilgariff on the Web as a corpus vs BNC

> The corpus resource for the 1990s was the BNC. Conceived in the 80s, completed in the mid 90s, it was hugely innovative and opened up myriad new research avenues for comparing different text types, sociolinguistics, empirical NLP, language teaching and lexicography.
>
> But now the web is with us, giving access to colossal quantities of text, of any number of varieties, at the click of a button, for free. While the BNC and other fixed corpora remain of huge value, it is the web that presents the most provocative questions about the nature of language. It also presents a convenient tool for handling and examining text.
>
> (Kilgariff 2001: 342)

The publication of Chomsky's revolutionary *Syntactic Structures* in 1957 heralded in a new era in linguistics, which gave rise to much debate on linguistics of the mentalist variety vs empirically based linguistics discussed in the following section.

2.2 Revolutionary and counter-revolutionary

Although Chomskyan linguistics dominated the linguistic scene in the 1950s and 1960s, those working in the empiricist tradition, such as Kucěra and Francis, mounted an attack with the compilation of computer corpora of naturally occurring data and gained considerable ground in this counter-revolution. Differences between Chomskyan linguistics and corpus linguistics are first presented as dichotomies below for ease of reference (Concept 2.1).

Concept 2.1 Differences between Chomskyan linguistics and corpus linguistics

Chomskyan linguistics	**Corpus linguistics**
Competence (internalised knowledge of a language, i.e. what *can* be said or written)	Performance (external evidence of language competence, i.e. what is *actually* said or written)

(*continued*)

Linguistic 'facts' accessed through introspection or intuitive means. Data are not 'objective', i.e. not verifiable	Linguistic 'facts' based on attested instances of authentic language. Data are observable and therefore verifiable
Structure: what is 'grammatical' or 'ungrammatical'? Judgements based on usually artificial, i.e. invented examples	Use: what are the 'degrees of grammaticality', and what is 'acceptable'? Judgements based on naturally occurring data
Potentially infinite number of examples, accommodating the concept of creativity	Finite number of examples which focus on formulaicity

However, as several corpus linguists have pointed out (cf. Leech 1992; McEnery and Wilson 2001; Stubbs 2001a), rather than viewing these two camps as diametrically opposed to each other, it may be more fruitful to focus on their respective insights into language and adopt a more accommodationist approach where the advantages and drawbacks of each are acknowledged.

Quote 2.4 Stubbs on Chomskyan linguistics and corpus linguistics

> When Chomskyan linguistics took a decisive step away from studying behaviour and its products, to studying the cognitive system which underlies behaviour, this led to the discovery of many new facts about language. Equally, when corpus linguistics took a decisive step towards the study of patterns across large text collections, this also led to the discovery of many new facts. The approaches are often seen as being in opposition, and the dualisms are perpetuated, but the long-term aim must be to integrate the insights from different approaches.
>
> (Stubbs 2001a: 242)

2.2.1 Competence vs performance

It is not an easy task to differentiate between linguistic competence and performance data, mainly on account of the fact that performance data can subsume other aspects. McEnery and Wilson mention short-term memory limitations and effects of drink on speech production as examples of non-linguistic factors. The picture is further clouded by the fact that Hymes (1971: 281), objecting to Chomsky's dualism, expanded the notion of competence to include, in addition to what is 'formally' possible, whether something is appropriate in relation to a context in which it is used and evaluated, which would overlap to a certain extent with McEnery and Wilson's view of performance data, i.e. 'usage on particular occasions'.

Concept 2.2 Distinction between competence and performance

> Competence is best described as our tacit, internalised knowledge of a language. Performance, on the other hand, is external evidence of language competence and its usage on particular occasions when, crucially, factors other than our linguistic competence may affect its form.
>
> (McEnery and Wilson 2001: 6)

It can be seen that both Hymes and McEnery and Wilson stress the importance of socio-pragmatic factors and psychological aspects in relation to performance data. And here we come up against a drawback of the corpus. Corpus-based research can only study performance data from the perspective of *product*; what is missing is the *process* aspect, accounting for socio-pragmatic competence and psychological inclinations of the interlocutors. Given that *process* factors, especially mental processes, are usually, of necessity, taken out of the equation in corpus data, then as Leech (1992) argues, there may be more of an alignment between competence and performance than is usually supposed.

Quote 2.5 Leech on competence and performance

> It can be argued that the putative gulf between competence and performance has been overemphasised, and that the affinity between (say) the grammar of a language as a mental construct and the grammar of a language as manifested in performance by the native speaker must be close, since the latter is a product of the former.
>
> (Leech 1992: 108)

2.2.2 Introspection and intuition vs attested instances of authentic language

Although introspection and intuition are often conflated, they do, in fact, represent slightly different mental processes. Introspection seems to be more to do with *retrieval* from one's mental lexicon, whereas intuition implies some sort of *interpretation* of data. In fact, Sinclair (1991: 39) seems to be hinting as much, when he states: 'It has been fashionable among grammarians to introspect and to trust their intuitions about structure …', with intuition being a higher-level cognitive skill than introspection. Wray (2005) identifies the following three levels of intuition, which all indicate this interpretative dimension: (i) ability to sense something is ungrammatical, (ii) ability to talk about grammar and (iii) ability to make judgements about the data we gather.

Reliability of intuition

In Section 1.3.2 I already touched on the issue of NS and NNS intuition regarding *interpretation* of corpus data. In this section I pick up this thread of intuition, looking at it from the perspective of using a corpus for *confirmation* or *denial* of our intuitions about language. While Wray seems to relate her levels of intuition mainly to the area of grammar, the corpus linguistics literature relates how intuition has been applied to frequency counts of various sorts, collocation, colligation and semantic prosodies. Interestingly, most of the reports in the literature point to how unreliable NS intuition can be (although see Owen 1996 for an instance of corpus evidence confirming his NS intuition that there seems to be something wrong with the phrase *Further experiments require to be done*).

Wray offers the following three reasons as to why intuition can be unreliable. Each of these reasons will, in turn, be discussed with reference to specific cases in the literature on the unreliability of intuition.

Quote 2.6 Wray on why intuition is unreliable

1. It is 'corrupted' by our psychological and cultural take on language;
2. Corpora are an insufficiently close match for the individual's linguistic experience;
3. Intuition offers a window on only partial, privileged information.

(Wray 2005: conference abstract)

1. It is 'corrupted' by our psychological and cultural take on language

Example 2.1 Faulty intuition concerning grammatical usage

Sinclair recounts the following example of how intuition based on phraseology overrides the common usage found in authentic corpus data:

> Overwhelmingly, *glad* is used predicatively, and in some complex constructions. However, many English speakers, when asked about the usage of this word, cite a phrase from the translation of the bible published in 1611 – *glad tidings of great joy* – that is still alive and well in the speech community. Apart from this relic, and a few minor phrases, *glad* will be found, on thousands and thousands of occasions, in predicative position. Without one of the tiny number of collocations like *tidings*, it will sound very odd indeed as an attributive adjective. … here is another place where our intuitions may appear to report falsely about the facts of language.
>
> (Sinclair 2004a: 45)

The above example concurs with Biber et al.'s (1998: 3) point that 'In many cases, humans tend to notice unusual occurrences more than typical occurrences, and therefore conclusions based on intuition can be unreliable'.

That many English speakers were found to cite *glad* in the context of a biblical phrase indicates the importance of cultural salience vis-à-vis corpus data (see Section 1.2.1). One could also argue that if this information was elicited around Christmas time, then this phrase is even more likely to be cited by speakers, possibly sparked by the memory of carol singing, adding a psycholinguistic construct to the data. One also wonders whether NNS would have had the same reaction to the item as the English speakers. In this case, NNS intuition might prove more reliable, if unencumbered by such cultural and psychological considerations.

2. Corpora are an insufficiently close match for the individual's linguistic experience

Example 2.2 Faulty intuition concerning semantic prosody

Channell (2000), cited in Hunston (2002: 21), notes that some semantic prosodies are not accessible via intuitive means. For example, she remarks that the phrase *par for the course* is recognised by native speakers as a comment that something frequently happens, but that speakers only realise that it is also used to evaluate an event negatively, when being informed so. Upon being told, they immediately recognise this prosody, an instance of what Louw refers to as 20/20 hindsight.

The example above suggests that there is a mismatch between an individual's usage and usage found in a corpus. Why this occurs could partly be accounted for by frequency. In fact, this phrase *par for the course* is relatively uncommon and only occurs 38 times in the BNC. Lack of familiarity with this phrase as it is fairly infrequent may well be a reason why speakers do not immediately connect it with a negative prosody.

3. Intuition offers a window on only partial, privileged information

Example 2.3 Faulty intuition concerning frequency counts

In a corpus study of metaphors in business journalism Partington (1998: 112–16) surmises that he expected to find lexis such as *war* and *battle* relating to the metaphor of BUSINESS AS WAR, given the prevailing outlook of business as a very competitive industry. However, such vocabulary was found to be far less frequent than that relating to BUSINESS IS FORECASTING OR GUESSING, thus denying his original intuitions about business writings.

Intuition may have failed in the above case due to the fact that the researcher is investigating a very specific sub-genre of business writing, i.e. business journalism. But his intuition about business writing could well be true for another sub-genre of business writing. Not being a member of a particular discourse community puts one at a disadvantage as one does not have the privilege of 'insider knowledge', so to speak. Similar to corpora, which are 'incomplete' as they can never fully represent all the facts about a language, intuition is also 'partial' and can never be fully relied on to make pre-emptive judgements. Moreover, unlike the two cases discussed above, here, the researcher is playing two roles – that of native speaker and that of linguistic researcher, making the data somewhat subjective and less verifiable than if they had been collected through informants (Aarts 2002a).

Following on from Wray's explanations as to why intuition is unreliable, it could be argued that alignment between a corpus and an individual's intuition would be likely when informants share the same set of usage norms. In support of this point, Stubbs (2001a: 71) remarks that: 'In many areas of semantics and pragmatics, intuitions are strong and stable, across all native speakers, whether linguistically naïve or trained, and must be given the status of data.' In fact, Partington (1998) comments on the overemphasis sometimes attributed to speakers on their lack of intuition regarding semantic prosodies. Citing the example of *crowd* and *mob* given in Stubbs, Partington points out that most people would be able to infer the different semantic prosodies of these items by intuitive means without having recourse to a corpus. Similarly, it is expected that most people would be able to intuit the general frequencies of core vs specific lexis, i.e. that *walk* is more common than *stroll* (Halliday 1993).

Given the fact that reports in the corpus linguistics literature on the role of intuition vs corpus data for interpreting and judging various aspects of language are rather sketchy with isolated accounts popping up here and there, this area would seem to be ripe for further research.

Status of corpus linguistics vs intuition, introspection and elicitation

Chomsky has been quite a vocal critic of corpus linguistics: 'It doesn't exist', stated in an interview with Aarts (cited in Aarts 2000: 5). For Chomsky, inquiry-based methods such as introspection are the only means of obtaining sufficient and relevant quantities of data. However, in an interview cited in McEnery and Wilson (2001: 11), Chomsky's response shows precisely why a corpus may be useful, and why NS intuition cannot always be relied upon. McEnery and Wilson (2001) point out that a check in the BNC shows that the verb *perform* can, indeed, be used with mass objects, e.g. *perform magic*.

Quote 2.7 Chomsky on native-speaker intuition

Chomsky:	The verb *perform* cannot be used with mass word objects: one can *perform a task* but one cannot *perform labour*.
Hatcher:	How do you know if you don't use a corpus and have not studied the verb *perform*?
Chomsky:	How do I know? Because I am a native speaker of the English language.

(Cited in McEnery and Wilson 2001: 11)

However, not all theoretical linguists are as dismissive of corpus linguistics as Chomsky. Chafe (1992) and Fillmore (1992, 2001) have also acknowledged the important role that corpora play. Likewise along with Johansson (1991: 313), 'The corpus remains *one* of the linguist's tools, to be used together with introspection and elicitation techniques', other leading linguists (e.g. Biber et al. 1998; Granger 2002) working within a more descriptive framework, have also emphasised that corpus linguistics is but one source of evidence for understanding language behaviour.

Quote 2.8 Fillmore on corpus linguistics and introspection

My peace-making position [in Fillmore 1992 – JA] was that one couldn't succeed in the language business without using both resources: any corpus offers riches that introspecting linguists will never come upon if left to their meditations; and at the same time, every native speaker has solid knowledge about facts of their language that no amount of corpus evidence, taken by itself, could support or contradict.

(Fillmore 2001: 1)

An important question arising from this three-pronged approach to investigating language behaviour concerns when one approach might be preferred over, or combined with, the others. For example, Quirk (1992) gives an example where elicitation tests were found to be more insightful than corpus data or introspection alone. Investigation of the alternatives *learned* and *learnt* in the Brown and LOB corpus showed a clear difference between American and British English, but it was the elicitations tests with separate groups of American and British students that revealed identity in bringing out a latent aspectual contrast between *learned* and *learnt*.

The data obtained from carefully constructed elicitation tests to uncover intuition-based judgements can also be usefully compared with corpus-based quantitative data. One such study (Takaie 2002) compared the responses of 100 native informants by different age groups and sexes on various constructions (e.g. *Just now* with the present perfect aspect) with comparable data from various large-scale general corpora. In some cases, there were disparities between the intuitive and corpus data, which the researcher sought to explain. More detailed studies along the lines of this one would be very useful as they may well provide a more 'rounded' analysis of language behaviour, and would help to clarify the role that intuition plays.

This subsection has highlighted the difficulties associated with intuition as a reliable indicator of many facets of language use, i.e. frequency, semantic prosody, pragmatic meaning etc., and examined the possible reasons for these. In addition, although it is well accepted that corpora are just one of the tools for investigation, there is the issue of which methodology, or combination thereof, might be the most appropriate to apply to a particular research study. Adducing and contrasting evidence from different data sources may help researchers to gain a better understanding of the status and value of different types of linguistic evidence.

2.2.3 Grammaticality vs acceptability

Parallelling Chomsky's position as a major opponent of corpus linguistics is Sampson (1996, 2005) in the other camp, who, in a very tightly argued treatise (2005) disputes Chomsky's theory of generative grammar, and hence the premises underlying Pinker's (1999) work. There are two major points to reconsider here in the light of naturally occurring data. One aspect Sampson queries is the fixedness of rules in generative grammar.

Quote 2.9 Sampson on 'fixedness' in generative grammar

> Steven Pinker comments that … the arguments of a verb, such as direct and indirect object, must occur closer to the verb than its 'adjuncts'…: 'we can say *gave the documents to the spy in a hotel*, but not *gave in a hotel the documents to the spy*'. Since arguments and adjuncts are related as essence and accident, respectively, it does not seem too surprising that what is essential is commonly placed closer than what is accidental to the thing they both apply to. But it came as news to me, when I read Pinker's book that English has a fixed rule about it. … It took very little time to disprove Pinker's statement from the one-million-word Brown Corpus of American English,

because the second sentence in that corpus (taken from a news story in the *Atlanta Constitution*) is a counterexample:

> The jury further said in term-end presentments that the City Executive Committee, which had overall charge of the election, 'deserves the praise and thanks of the City of Atlanta' for the manner in which the election was conducted.

In this case, the reason why the argument is further from the verb than the adjunct is because the argument is much longer.

(Sampson 2005: 163)

Second, is the fact that the grammatical/unacceptable divide may not have such clearly defined boundaries, as shown by another of Sampson's counter-examples.

Quote 2.10 Sampson on 'acceptability' in generative grammar

… there was general agreement that multiple central embedding in some sense of the concept does not happen in human languages … For generative grammarians, who laid weight on the idea that grammatical rules are recursive, there was a difficulty in accounting for rules which could apparently apply once but could not reapply to their own outputs: they solved the problem by arguing that multiple central embeddings are perfectly grammatical in themselves, but are rendered 'unacceptable' by falling foul of psychological language-processing mechanisms which are independent of the rules of grammar but which, together with the latter, jointly determine what utterances people can produce and understand (Miller and Chomsky 1963: 471).

(Sampson 1996: 16)

The fact is that we *do* cope with it [multiple central embedding], frequently. For instance, here is a sentence from a published academic article about linguistics written by a Canadian:

> The only thing that the words that can lose -d have in common is, apparently, that they are all quite common words.

This is a multiple central embedding in precisely the same sense as Pinker's sentence about the malt and the rat [*The malt that the rat that the cat killed ate lay in the house*].

(Sampson 2005: 160)

Sampson's examples serve to illustrate how the use of invented, artificial language and naturally occurring data can lead to quite different conclusions on what is considered grammatical or acceptable.

As far as acceptability is concerned, decisions regarding what is 'acceptable' are usually far more contentious than decisions regarding what is 'grammatical' (Carter 1987). In the area of collocations, Carter suggests employing 'techniques of informant analysis in which the intersubjective intuitions of a group of native-language speakers are statistically measured and a line drawn between what can generally be allowed and what cannot' (p. 55). However, there are still discrepancies to be resolved between what native speakers say and what a large-scale corpus throws up.

Concept 2.3 NS judgements on acceptability vs evidence from corpus data

Nesselhauf (2003: 241) asked four NS informants to judge the acceptability of *run the danger*. All 'considered it wrong or of doubtful acceptability', preferring the more common *run the risk*. According to Nesselhauf, in the BNC the collocation *run the danger* cropped up 11 times in 10 different texts, with about 330 occurrences of *run the risk* found in about 310 texts. Here, we are presented with an instance where the corpus data do not seem to fully support native speaker intuition. The examples, admittedly very few, raise a question mark over the non-acceptability judgement of native speakers, thus hinting that NS intuition may not be infallible.

For further investigation of this problematic collocation, I conducted a Google search on it and found a BBC news report article, where it was used in the phrase 'they run the danger of contracting avian flu'. Such evidence highlights the fact that collocations are, to a great extent, context-dependent and that 'core' collocations may be slightly modified, as in the BBC news report, for rhetorical purpose. 'Danger' has a greater sense of impending threat than 'risk'. As Partington (1998: 19) cautions:

> … in the light of the dependency of collocational acceptability on register and style, such a demarcation [between acceptable/unacceptable] would have to be conducted with great care, making sure enough context was supplied to render informants' decisions meaningful.

The above example demonstrates the importance of contextual features in making acceptability judgements and that raw corpus data, by themselves, may not be sufficient unless interpreted with recourse to the situational and communicative context.

2.2.4 Creativity vs formulaicity

Chomsky's position is that natural language is non-finite, thus allowing a user to create a potentially infinite number of new utterances, which may never have been seen or heard before. On the other hand, corpus linguists tend to focus on the routinised expressions uncovered in naturally occurring data; but however large a corpus, it is always a finite entity.

This then brings us to the question of whether creativity can be accommodated within the parameters of a phraseological approach to language. For Wray, pattern grammar lends itself very well to creativity (see Quote 2.11). Likewise, Partington (1998: 121) states that 'much of what is seen as creative use of language is not invented "out of the blue", but is an imaginative *reworking* of the usual'. This perspective is also in line with other linguists' views (cf. Weinert 1995) for whom the distinction between creativity and formulaicity is not a dichotomy, but more of a continuum.

Quote 2.11 Wray on pattern grammar and creativity

> Pattern Grammar has no difficulty with accommodating creativity, which simply, is the use of an unaccustomed word with an established pattern, though significantly usually one which it is fairly easy to link semantically or pragmatically (e.g. metaphorically) with the customary lexical associates of that pattern since the effect of the new association relies on the resonances it creates with what *might* have been said.
>
> In this respect, Pattern Grammar nicely codifies the semi-fixed frames which are a major feature of formulaicity in language, and accommodates with ease the relationship between routine and creativity.
>
> Wray (2002: 275)

Creative language, it should be emphasised, is not confined to literary works (see Section 6.4.2), but occurs across a wide range of genres. Carter (2004) and Carter and McCarthy (2004) illustrate that creative use of language is much more pervasive than one would expect in everyday conversational English (see Section 3.4 for further details on Carter and McCarthy's sociolinguistic approach to corpus analysis).

Example 2.4 Creative language use in the 5-million-word CANCODE corpus

<S 03>	I reckon it looks better like that
<S 02>	And it was another bit as well, another dangly bit
<S 03>	What, attached to

(*continued*)

<S 02> The top bit
<S 03> That one
<S 02> Yeah. So it was even
<S 03> Mobile earrings
<S 01> I like it like that. It looks better like that

Carter and McCarthy comment thus on the striking use of metaphorical wordplay of *mobile* with *earrings*:

> There is a pun on the meaning of 'mobile' (with its semantics of movement) and the fixture of a *mobile* – meaning either a brightly coloured dangling object which is normally placed over a child's bed or cot to provide distraction or entertainment, or else which is a piece of moving art.
>
> (Carter and McCarthy 2004: 64)

Creative use of language is also to be found in journalistic writing, where metaphorical systems are reworked (Deignan 2005).

Example 2.5 Creative language use in newspaper text extracted from BofE

(9) Election fever reaches its climax tomorrow – with the White House candidates *hurtling almost neck-and-neck towards the finish*.

In the above example from Deignan (2005: 30), the conceptual metaphor, WINNING IS A RACE, has been extended by the inclusion of non-conventional linguistic metaphors (i.e. *hurtling* and *finish*). These have the effect of foregrounding the metaphorical expression, *neck-and-neck*, which would usually be in its unmarked conventionalised form and therefore go unnoticed. The juxtaposition of *hurtling* and *finish* with *neck-and-neck* creates a novel expression. Interestingly, Deignan remarks that as well as the metaphorical interpretations, the literal meanings of these metaphors will probably be evoked, with possible humorous effects. This example also serves to illustrate that it may not always be possible to interpret an expression as absolutely either literal or metaphorical, as one meaning may shade into the other (see Sinclair's example of *budge* in Section 1.2.2).

Creative use of language can also result from the manipulation of both collocational and colligational aspects simultaneously.

Example 2.6 Manipulation of collocation and colligation for creative effect

At a TESOL conference a few years ago, Geoffrey Pullum gave a presentation on the conceptual framework of *The Cambridge Grammar of the English Language* (Huddleston and Pullum 2002). Although this grammar is not corpus-based, Geoffrey Pullum did make the point with illustrative examples (one being a sentence containing *mere*) that some adjectives are more likely to occur in attributive than predicate position. As Chair of the session, Alan Davies thanked the speaker and concluded the session with the memorable ending '… and his talk was not mere'. Not only is there a play on words here with a somewhat unusual collocation of *talk* with *mere* (in the BNC there are only two instances of *mere* collocating with *talk* and these occur with the uncountable form, a slightly different meaning from the countable one used), but Alan Davies also manipulated colligational features as he transposed an attributive adjective into a predicative one.

The above examples would seem to fall into a category that Hoey (2005: 153) sees as existing between the Chomskyan sense of creativity and the literary sense of creativity '… sentences that make no claim to be literary but which surprise us in some way, either because they draw attention to themselves by their clever wording or because they are momentarily hard to process or make us aware that they are indeed made of language'. Creativity can also involve reworking of grammar, but this usually involves creativity of the literary kind. For instance, Wray (2002: 12) cites a line from an e.e. cummings poem: *he sang his didn't he danced his did* to show how grammar can be manipulated through the changing of word classes.

The examples above demonstrate that creativity can indeed be encompassed within the phraseological approach, involving quite a complex manipulation and reworking of phraseological and grammatical features in varying degrees along the formulaic–creative continuum, manifested in a wide range of genres.

This chapter has sought to lay out the historical and conceptual background of corpus linguistics against a backdrop of empiricism stemming from the experimental approach to scientific enquiry with the establishment of the Royal Society in the seventeenth century, and Chomsky's revolutionary mentalist approach to linguistics. At the same time, the discussion on the dualisms between Chomskyan linguistics and corpus linguistics has also sought to reconcile the differences between the two approaches, leading to a more accommodationist perspective on some aspects of what has formerly been perceived as a rationalist–empiricist divide.

Further reading

Andor, J. (2004) The master and his performance: an interview with Noam Chomsky. *Intercultural Pragmatics*, 1(1): 93–111. This paper is a transcript of an interview with Noam Chomsky in which he reiterates his misgivings about corpus linguistics.

Gilquin, G. and Gries, S. Th. (2009) Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5 (1): 1–26. Using a sample of recent studies, this paper shows how psycholinguists combine corpus and experimental data.

de Mönnink, I. (1997) Using corpus and experimental data: a multimethod approach. In M. Ljung (ed.) *Corpus-Based Studies in English*, pp. 227–44. Amsterdam: Rodopi. This empirical paper shows how experiment data in the form of elicitation tests can be used to supplement corpus findings with a view to deciding which constructions should be incorporated in the descriptive grammar.

Tognini Bonelli, E. and Sinclair, J. McH. (2006) Corpora. In K. Brown (ed.) *Encyclopedia of Language and Linguistics*, 2nd edn, pp. 216–19. Amsterdam: Elsevier. This article presents a concise overview of the history of corpus linguistics.

3
Approaches to Corpus Linguistics

This chapter presents an overview of the five main schools of corpus-based approaches to linguistic analysis:

- Neo-Firthian approach
- Probabilistic approach to grammar
- Systemic-functional grammar approach
- Multidimensional approach
- Sociolinguistic approach

In much of the discussion in the following sections, I focus on the overlap and differences in the theoretical underpinning and methodological procedures of the five approaches with reference to major works in the respective schools.

3.1 Main British traditions in corpus linguistics: probabilistic approach to grammar and neo-Firthian approach

There are generally considered to be two main British traditions in corpus linguistics. One broadly derives from the probabilistic approach to grammar adopted on the Survey of English Usage (SEU) project started by Jan Svartvik in the early 1960s at University College London, involving Quirk, Greenbaum and also Leech, who later moved to Lancaster and instigated the compilation of the Lancaster-Oslo-Bergen (LOB) corpus on the same principles as the SEU. The other, often referred to as the neo-Firthian approach, has as its main proponents Halliday, Sinclair, Stubbs, Hoey and Hunston. The former approach uses corpus data for the study of grammatical categories such as those providing the framework for *A Comprehensive Grammar of the English Language* (Quirk et al. 1985). The latter approach, meanwhile, derives its theoretical framework for analysis of corpus data from Firth's 'contextual theory of meaning' and concept

of collocation (see Section 1.2.2), with the University of Birmingham as the main centre for research connected with the Firthian tradition of linguistics.

The differences in the theoretical stances of these two strands of corpus research are reflected in their procedures of corpus assembly and treatment of the corpus data, tabulated in Table 3.1.

Concept 3.1

Table 3.1 Probabilistic vs neo-Firthian approach

Probabilistic approach to grammar	**Neo-Firthian tradition**
Sampling of text extracts of different genres	Whole texts
Set size of corpus	Open-ended corpus ('monitor' corpus)
Tagged using probabilistic grammar methods	Untagged to show probabilistic tendencies

Sinclair has been a keen advocate of compiling corpora of whole texts rather than text segments, having an open-ended corpus which can be added to, and keeping the text as unprocessed and clean of any tags or other codes in order to accommodate the somewhat amorphous nature (in the sense of defying 'neat categorisation') of corpus data. Sinclair justifies the use of whole documents by noting the drawbacks of text extracts. First, there may be marked differences between different sections of a textbook as not many features of a book-length text are evenly distributed throughout. Another point is that an even sample size of 2000-word text extracts may not accurately reflect the overall size of the documents from which they are drawn. Sinclair's notion of an ever-evolving, open-ended 'monitor' corpus is an intriguing one.

Quote 3.1 Sinclair on a 'monitor' corpus

> It is now possible to create a new kind of corpus, one which has no final extent because, like the language itself, it keeps on developing. Most of the material will come in from machine-readable sources, and it will be examined for the purposes of making routine records. Gradually, it will get too large for any practicable handling, and will be effectively discarded. The focus of attention will be on what information can be gleaned from the text as it passes through a set of filters which will be designed to reflect the concerns of researchers.
>
> (Sinclair 1991: 25)

As the thrust of the research at Birmingham is on lexicography and phraseology for the compilation of general-purpose dictionaries and grammars (see Section 6.7), this kind of corpus resource would seem to be ideally suited for creating and recreating a 'state of the language' for such types of applications reflecting the changing nature of language. However, for other purposes where research questions have been defined a priori to the compilation, specialised corpora of a finite size are perfectly adequate.

In the probabilistic approach to grammar associated with the SEU project, automatic taggers are used to assign grammatical category labels to natural language text based on the probabilities of possible adjacent tags. Such taggers have difficulty disambiguating a few categories such as adjuncts vs prepositional phrases. Aside from such syntactic ambiguities, though, the main reason Sinclair argues for a corpus to be untagged is on account of the probabilistic tendencies of language. As Stubbs (1996: 40) points out, 'Any grammatical structure restricts the lexis that occurs in it; and conversely, any lexical item can be specified in terms of the structures in which it occurs', which is the foundation of the pattern grammar approach expounded in Hunston and Francis (2000). For Sinclair, working with a tagged corpus would veil the interdependency existing between grammar and lexis by setting up artificially imposed categories on the language, and disallow language to speak for itself, as it were. Sinclair's exhortation is to 'trust the text' and allow the corpus data, devoid of any codes, to suggest patterns which may throw up new insights into language hitherto unnoticed. This philosophy is at the heart of the 'corpus-driven' approach (Tognini Bonelli 2001) which contrasts with the more 'corpus-based' approach associated with the SEU and LOB schools where the corpus data are defined in terms of a particular theory of grammar, which is itself reflected by the annotation system used (see Section 4.1 for a discussion of the terms 'corpus-driven' and 'corpus-based').

3.1.1 Methodological issues in the neo-Firthian approach

One issue arising from the 'corpus-driven' approach is whether the starting point is with the pattern grammar or lexis.

Quote 3.2 Pattern vs lexis as point of entry for investigation

> Sinclair's work is based on possibly two contradictory methodologies. One involves the researcher painstakingly investigating the phraseology of one lexical item after another. The other involves the use of a computer to list the most frequently-occurring word sequences. Francis faces the same problem of whether to take the lexical item as the starting point or whether
>
> (*continued*)

> to take the patterns as the starting point. She investigates the adjective, *possible*, for example, and notes that it occurs with an unusually wide range of patterns, each of which it shares with other adjectives. On the other hand, she investigates patterns such as the appositive that-clause, and lists the nouns which share that pattern.
>
> For her study, Francis advocates 'moving from environment to item and back to environment again' (Francis 1993: 146). For example, her study of the pattern (environment) '*it* + link verb + adjective + that-clause' leads to an interest in an item, *possible*, and then to an investigation of the other patterns (environments) in which that adjective occurs. Similarly, she suggests that the investigation of appositive that-clauses 'could profitably lead to an exploration of the grammar of some of the more frequent noun-heads, of which *fact* and *reason*, for example, promise rich findings' (Francis 1993: 155).
>
> (Hunston and Francis 2000: 31)

Which methodology is used as an entry point would very much seem to depend on the purpose and scope of the investigation. For compiling a comprehensive lexico-grammar of the English language it may be best to start with the pattern and identify all the words that have a particular pattern. However, as Hunston and Francis (2000) point out, it would be rather unwieldy, for example, to list all the nouns with the pattern N *of* N.

Neither would this system capture relations existing between patterns, such as the following with introductory *it*, which all perform the same function (p. 35):

***it* V n to-inf**	e.g.	*It hurts me to think of that*	verb pattern
***it* v-link N to-inf**	e.g.	*It would be a shame to lose touch*	noun pattern
***it* v-link ADJ to-inf**	e.g.	*It was terrible to see his face*	adj. pattern

However, where the aim is to examine the lexico-grammar in a specialised corpus, it may be more opportune to start with the lexis, for as we have seen in Section 1.2.1, Sinclair (2005) notes that 'the characteristic vocabulary of the special area is prominently featured in the frequency lists'. Lexis is also the starting point for Hanks' (2009) *Pattern Dictionary*, which aims to capture the phraseology of all the patterns of a particular verb (see Section 6.7.5).

3.1.2 Identification of phraseological units

Another methodologically related issue that has occupied linguists concerns how phraseological units are to be identified. Various studies, which take either a corpus-based or psycholinguistic approach from different perspectives within these two broad approaches, are discussed below. As mentioned in Section 1.2,

there abounds a plethora of terms for this phenomenon. As corpus linguists tend to use expressions containing the string 'phrase', and psycholinguists use expressions with the string 'formula', these are the terms I shall, in general, adopt in the discussion below.

Corpus-based approaches to phraseology

Corpus-based approaches to identifying phraseological units tend to be pragmatic, relying as they do on various software packages to obtain statistical information on recurrent syntagmatic patterns. But what exactly constitutes a phraseological unit is not always easy to determine.

Quote 3.3 Altenberg on types of phraseology

> Phraseology is a fuzzy part of language. Although most of us would agree that it embraces the conventional rather than the productive or rule-governed side of language, involving various kinds of composite units and 'pre-patterned' expressions such as idioms, fixed phrases, and collocations, we find it difficult to delimit the area and classify the different types involved. Indeed, as Pawley and Syder (1983) and others have pointed out, the existence of a large number of more or less prefabricated expressions in language blurs the distinction between lexicon and grammar and strongly suggests that 'lexicalisation and productivity are matters of degree' rather than a clear-cut dichotomy. This state of affairs creates problems of description for both the empirical and theoretical linguist, at the same time as it provides a challenge to anyone who wants to get a better understanding of language and language use.
>
> (Altenberg 1998: 101)

One classification which aims for a comprehensive coverage is that by Erman and Warren (2000). Acknowledging the difficulty of identifying what they refer to as prefabs, Erman and Warren have come up with the following four categories: lexical, defined according to notional categories such as places and positions, e.g. *here and there*, *to the right*; grammatical, one subcategory of which denotes intensification, e.g. *more or less*, *much less*; pragmatic prefabs with a functional orientation including hedges, e.g. *sort of*, and attitudinal markers, e.g. *I must say*, and a somewhat puzzling category termed reducibles for contractions, e.g. *I'm*, *don't*, which raises the question of what constitutes a phraseological unit as usually contractions are treated as a single word (cf. Biber et al. 1999).

Another classificatory framework is that by Moon (1998) who first reviews various phraseological models (lexicalist, syntactic, functional and

lexicographical), then proposes her own categorisation for what she terms FEIs, fixed expressions and idioms. Based on deficiencies in existing typological models, Moon developed three macrocategories, namely anomalous collocations, formulae and metaphor: 'This typology essentially involved identifying the reason or reasons why each potential FEI might be regarded *lexicographically* as a holistic unit: that is, whether the string is problematic and anomalous on grounds of lexicogrammar, pragmatics, or semantics' (Moon ibid.: 19).

A smaller-scale study is that by Altenberg (1998), one of the first linguists to explore phraseology in corpora. Altenberg adopts a two-tier classification, with a structural and then functional categorisation, for some of the recurrent word combinations in the London-Lund Corpus (LLC) of Spoken English, as given in Table 3.2.

Altenberg's detailed study also shows that sequences often have a core, with optional extensions (e.g. [*oh/yes*] *I see*), and that sequences may overlap, sometimes interrupted by non-formulaic language in accordance with Sinclair's 'open choice' principle, or follow a 'stitching' model of discourse production as an alternative to a more rule-governed one (e.g. *but I mean are you going to do it at all*; *because you see I don't want to see you at the moment*). In fact, there seems to be general agreement among corpus linguists that language is a mixture of formulaic ('idiom principle') and non-formulaic ('open choice principle') elements. Both Wray and Sinclair are of the view that formulaicity assumes central position and that the open choice principle operates only when formulaic language is insufficient.

The identification of phraseological units based on frequency counts is also problematic, being dependent on the type of software used. The cluster facility in WordSmith Tools (Scott 1999) allows only strings which recur in identical form to be extracted. Such software would not be able to capture the different degrees of variability within recurrent sequences, as permitted by software such

Example 3.1 Classification of recurrent word combinations

Table 3.2 Recurrent types of dependent clauses

Functional type	**Example**	***n***
Comment clauses	*as it were*	23
	I should think	20
	as you know	18
	as I say	12
Indirect conditions	*if I may*	12
	if you like	11
Apposition marker	*that is to say*	11

(Altenberg 1998: 109)

as Kwikfinder (Fletcher 2002) and ConcGram (Greaves 2009). Variability could also be compromised by the span which is set for the node word. As Stubbs (1995b) points out, while spans of 2:2 or 3:3 are often used, a window of 3:3 would capture examples such as *the cause of the trouble*, but not *the cause of all the trouble*, where *cause* is the node word.

There thus exists a great deal of diversity regarding phraseological units in terms of length, boundaries, variation allowable and the methodologies for their identification in corpora. Given these different definitions of phraseological units, it is not surprising that there are very diverging figures given in the literature (5–80%) for the incidence of such units (Lindquist, Corpus Linguistics Discussion list 14.3.06). However, regardless of the types of classifications for phraseological units and whether these are computed wholly automatically or using a combination of semi-automatic and manual procedures, what all the corpus-based studies cited above have in common is that they make reference to frequency counts, to a greater or lesser degree.

Reliance on frequency counts is not unproblematic, though, one main reason being that the cut-off points chosen, for both recurrent combinations and their overall frequency, are, by necessity, often arbitrary.

Quote 3.4 Altenberg on frequency counts

> … the sheer bulk of the material makes some sort of selection necessary. For practical reasons, I will therefore limit my examination to word-combinations consisting of at least three words occurring at least ten times in the corpus. These limitations are to a large extent arbitrary. Neither length nor frequency is a criterion of phraseological status, but the frequency threshold gives at least some guarantee that the selected word-combinations have some currency in spoken discourse and that they are of some interest from that point of view. The length restriction was chosen partly to reduce the number of fragmentary sequences, but mainly to reduce the material to a manageable size.
>
> (Altenberg 1998: 102)

Wray (2002: 31) also pinpoints other disadvantages of relying on frequency-based counts. First is that frequency counts will not differentiate between those sequences which are formulaic under some circumstances but not others. For example, *keep your hair on* is formulaic when it means 'calm down', but not when it means 'don't remove your wig', which cannot be disambiguated by present software, but only by contextual and pragmatic cues. Secondly, in common with Altenberg, Wray mentions that frequent strings are not necessarily an indicator of prefabricated language (see also Wray 2008).

Quote 3.5 Wray on the relationship between frequency and formulaicity

> … just as there is evidence that a string generally agreed to be formulaic may or may not have a high frequency in even the largest of corpora, so it is also not possible to assert that all frequent strings are prefabricated. It can, it is true, be argued on theoretical grounds that, if a string is required regularly, it is likely to be stored whole for easier access (e.g. Becker 1975; Langacker 1986: 19–21), but it does not have to be. In order to distinguish between frequent strings that were and were not prefabricated, we should therefore need an independent set of supplementary criteria.
>
> (Wray 2002: 31)

Having reviewed corpus-based approaches to phraseology together with their concomitant drawbacks, I will now briefly examine what psycholinguistic approaches have to offer in this complex area and discuss 'the set of supplementary criteria' that psycholinguists put forward for delineating recurrent sequences.

Psycholinguistic approaches to phraseology

In contrast to the identification of phraseological units, which rely heavily on frequency counts and which are usually defined in terms of the lexico-grammar or functional units of meaning in corpus linguistics, such type of sequences are characterised by psycholinguists through their socio-psychological function rather than in formal or semantic terms. As individuals construct their own socio-psychological reality, this means that formulaic sequences are not seen as a fixed store of items, but can vary from one individual to another. One point on which both corpus linguists and psycholinguists seem to agree, though, is that the starting point for generating these sequences is not the language grammar, but resides in their meaning potential.

Quote 3.6 Wray on definition of formulaic language

> A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.
>
> (Wray 2002: 6)

Various experimental studies, involving pauses, eye movements, self-paced reading aloud and dictation tasks have been conducted to uncover the composition and boundaries of formulaic sequences with tentatively positive

results. For example, Wray (2004) found pauses to be much less frequent inside formulaic sequences than outside them, and Erman (2007) noted that pausing was rare between the component parts of prefabs, thus concluding that prefabs are stored in long-term memory as units (see Schmitt 2004, for further discussion on this issue).

But do the frequency counts in corpus-based approaches used as indicators of formulaicity actually reflect the holistic and mental storage and processing of sequences (Butler 2004a)? The answer to Butler's query is not affirmed in one study but partly in others. A research study conducted by Schmitt et al. (2004) indicated that corpus-derived recurrent clusters (e.g. *in the number of*; *as a matter of fact*) were not accurately reproduced from mental storage by task-takers in a dictation activity. A more promising study is that by Ellis et al. (2009), whose research investigated the psycholinguistic reality in language users of the phenomena of collocation and semantic prosody using corpus analysis data. Two different approaches to collocation, statistical and textual, were discussed in Section 1.2.2. Ellis et al.'s data present empirical evidence for a third approach, psychological, i.e. associative, as outlined in Concept 3.2.

Concept 3.2 Experiments to test whether word recognition is sensitive to collocational frequency and semantic prosody

Experiment 1
This experiment tested whether word recognition was sensitive to collocational frequency. The collocations tested were taken from Kennedy's (2003) study of amplifier patterns in the BNC (e.g. *absolutely diabolical, entirely blameless*) and Kennedy's (2005) research on collocational patterning with high-frequency lexical verbs using sets of semantically related verbs (e.g. *start, begin*). In this language processing task participants were required to judge a pair of letter strings as either both words or not (e.g. *cause problems* and *phrup problems*). Results showed that the higher the collocational strength of an item, the faster the participants were able to recognise that both of them were words, thus showing that language processing is clearly sensitive to patterns of usage of particular collocations verifying the psycholinguistic reality of collocations.

Experiment 2
This experiment tested whether word recognition was sensitive to semantic prosody, based on Kjellmer's (2005) corpus data. Verbs judged to have strong positive or negative semantic prosody were selected for the study. A similar experimental paradigm was adopted to that in the first experiment. However, semantic prosody was not shown to have an effect on word recognition fluency and lexical access.

(*continued*)

Ellis et al. thus conclude that: 'It appears that fluent lexical access is due to memory for particular lexical association – there are no top-down semantic generalizations upon this level of processing' (2009: 106).

In view of the fact that semantic prosody is quite a contentious concept (see Section 1.2.2), it may therefore not be so surprising that experiment 2 did not verify its psycholinguistic reality in cognitive processing.

From the paucity of research examining congruence between corpus-derived and mental storage and retrieval of formulaic sequences, it seems that this is an area which deserves more attention to help clarify the nature and boundaries of such formulaic language.

3.1.3 Relationship between phraseological units and cognitive linguistics

Prototypical phraseological units, i.e. formulaic sequences associated with Sinclair's 'idiom principle', have not only been discussed from a frequency-based vs psycholinguistic perspective, as shown in the previous section, but also in relation to cognitive linguistics, a usage-based model developed by Langacker (1988, 2000) among others. The usage-based model of language cognition views meaning as paramount and for this reason is closely related to semantics. Moreover, meaning is considered to be experientially based and entrenched in actual language use, hence its nomenclature as usage-based.

Concept 3.3 Cognitive linguistics

In cognitivism, linguistic representations are grounded in usage events, with context playing an important role in the operation of the linguistic system. The linguistic system is built up from a cognitive representation of individual expressions with a gradual abstraction of more general representations from the repetition of similar instances of contextualised use. The abstracted patterns, often referred to as schemas, and their instantiations coexist in the grammar:

> The coexistence in the grammar of the schema and instantiations affords the speaker alternate ways of accessing a complex but regular expression with unit status; it can simply be activated directly, or else the speaker can employ the schema to compute it. (Langacker 1988: 129)

However unlike in the psycholinguistic approaches to phraseology discussed in Section 3.1.2, linguistic representations are not stored as fixed phrases, but are viewed as cognitive routines existing as activation patterns

in the processing of the linguistic system. Frequency plays a crucial role in cognitive routinisation; the higher the frequency of a particular pattern or its instantiation, the greater the degree of entrenchment in the linguistic system leading to the notion of linguistic representations as emergent in the system. These cognitive routines play a double role in the system, not only emerging from the linguistic system but also feeding back into it.

Several linguists have pointed out affinities in various areas between cognitive linguistics and corpus linguistics. Schönefeld (1999: 152) speaks of 'a direct correspondence between the corpus linguistic assumption of an "idiom principle" besides an "open-choice principle" and the cognitivist assumption of the cognitive representation of individual expressions or instantiations besides that of an abstracted pattern or schema'. Likewise, albeit from a slightly different perspective, Mukherjee (2004a: 96) recognises a close correspondence between cognitive linguistics and corpus linguistics. Stressing the notion that this usage-based model encompasses the concept of 'communicative competence', which in turn emphasises 'the ability to use linguistic forms and structures idiomatically (e.g. in terms of frequently co-occurring forms) and appropriately (e.g. in terms of pragmatic principles)', Mukherjee makes the case that a usage-based model can be derived from frequencies in corpus-based analyses of actual language use.

Several empiricist studies (cf. Gilquin 2003; Gries 2006) have sought to integrate corpus linguistics and cognitive linguistics through analysis of the data within a theory of frame semantics (Fillmore 2006), a facet of cognitive linguistics with its focus on usage-based meaning. It should be noted that corpus-based frame semantic analyses, such as the one by Gilquin below, are also based on case grammar theory (cf. Chafe 1970; Fillmore 1968), from which frame semantics originally developed. For example, Fillmore (1968) noted that impact verbs such as *hit, push, shove* can receive two different case analyses. The sentence *John hit the ball* can simply be read as *hit = X hits Y*. Alternatively, Fillmore postulated that *hit = X causes-Y-to-move-by-hitting it*, with the lexical verb indicating the cause of the motion.

Concept 3.4 Frame semantics

The main concept behind frame semantics is that the meaning of a word cannot be understood without activating the categories of real-world experience that relate to that word. For meanings to be instantiated speakers must have first-hand knowledge of a word's conceptual structure.

Gilquin's (2003, 2006) corpus-based study of causative verbs illustrates the advantages that insights from cognitive semantics can bring to corpus research.

Concept 3.5 Advantages of a frame semantic analysis for corpus data

The basic frame elements making up the causation frame are CAUSER, CAUSEE and EFFECT, with PATIENT sometimes included, as exemplified in Gilquin (2003):

The explosion	***caused***	*the temperature*	*to rise*.	
CAUSER		CAUSEE	EFFECT	
She	***had***	*them*	*send*	*her mail*.
CAUSER		CAUSEE	EFFECT	PATIENT

Gilquin (2003: 128) points out that a frame semantic analysis has advantages over a traditional syntactic description as it can account for the conceptual elements above, which may not be apparent in a traditional description.

(10) I had the student sit down.
(11) The technician had the video working.
(12) The emperor had the slave imprisoned.

Gilquin points out that, according to *A Communicative Grammar of English* (Leech and Svartvik 1994), all the above sentences would be categorised as consisting of a subject + causative verb + object + nonfinite complement. However, a frame semantic analysis would differentiate between (10) and (11) analysed as CAUSER, CAUSEE and EFFECT, and (12), which contains a PATIENT but whose CAUSEE is not expressed in this case, but would be in a sentence such as 'The emperor had *his guards* imprison the slave'.

Also working within a framework of frame semantics, Barlow (1996) investigated instances of actual usage of reflexives through the analysis of their different patterns, in which grammatical units are viewed in terms of form–meaning pairings. Corpus data revealed that, contrary to intuition-based studies which focus on constructions such as *He cut himself*, the most common schema in his corpus are those which involve self-observation with some degree of distancing, e.g. *I see myself as …*; other patterns in the corpus include those with a discourse-based function, e.g. *and yourself?*, which could be uttered by a waiter taking orders in a restaurant. Likewise, Gilquin (2006) in her study of causation found a discrepancy between the three prototypical theoretical constructs of causation identified in the literature and their occurrences in corpus data.

The above studies show that frame semantic analyses, which are ignored in traditional grammars, do indeed have a place to play in enhancing corpus analyses. However, more investigations on, or clarification of, the notion of prototypicality are needed, as called for by Gilquin (2006). Although Mukherjee (2004a: 96) asserts the strength of Schmid's (2000: 39) Corpus-to-Cognition Principle, i.e. 'frequency in text instantiates entrenchment in the cognitive system', several studies discussed in this section show this not always to be the case, with a gulf existing between prototypicality and frequency of occurrence.

3.2 Systemic-functional grammar (SFG) approach

The application of corpus-based linguistic techniques to systemic-functional linguistics is probably best discussed in relation to Sinclair's approach as major publications in the field of corpus linguistics often make comparisons between the two.

3.2.1 SFG approach vs phraseological approach

Both the SFG and phraseological approach view lexis and grammar as interconnected and as one and the same phenomenon. However, in the SFG approach paradigmatic choices are modelled according to a hierarchical structure of system networks, whereas in the phraseological approach, as already noted, language is seen as a linear, syntagmatic sequence.

Quote 3.7 Differences between Sinclair's and Halliday's approach

> Sinclair is by nature a lexicographer, whose aim is to construct the grammar out of the dictionary. I am, by nature, a grammarian and my aim (the grammarian's dream, as I put it in 1961) is to build the dictionary out of the grammar. … The point is that grammar and vocabulary are not two different things; they are the same thing seen by different observers. There is only one phenomenon here, not two. But it is spread along a continuum.
>
> (Halliday 1992: 63)

In the SFG approach, the relationship existing between the grammatical and lexical ends of the spectrum is one of delicacy with 'lexis as most delicate grammar'. For example, at the grammatical end we have a choice of various types of process verbs (relational, verbal, mental, material, behavioural, existential). Moving towards a greater degree of delicacy, one subset of a particular process could be chosen, e.g. *changing* from the subsets *happening*, *changing* and *acting* for material processes. At the most delicate level, a specific choice of lexical item would be made, for example whether for the process of *changing*, the

verb, *create*, *cause* or *result in* should be chosen which may well be dependent on the semantic prosody of a particular collocation. In contrast, in the phraseological approach the starting point would usually be with the lexical item, e.g. *cause* which would have its own collocational behaviour and grammar, as uncovered by evidence from a corpus. Lexis is thus the terminal point in the SFG system, but usually the starting point in the phraseological approach.

In Halliday's system networks for meaning making, choices have to be made at various points and it could well be that the material process of 'cause' is realised by a paratactic or hypotactic conjunction (e.g. *so*, *because* or *as*) in the form of a clause nexus rather than lexically. (Halliday (2004) points out that in spoken language such semantic relations are more often grammaticalised than realised through nominals.) While the phraseological approach views lexis as driving the grammar, it does not exclude the grammar as a point of entry for corpus investigation (see Section 3.1.1), but grammatical categories do not lend themselves so easily to searching. For instance, it would be much easier to trawl the corpus for explicit causative verbs than to extract occurrences of occasions when 'and' has a resultative function.

Hunston has also raised the issue of whether the SFG system network can accommodate the choices between subtle variations in a phraseological unit. Citing several studies comparing the SFG with the phraseological approach, Butler (1998, 2003a, 2004b) shows that they are in fact reconcilable to a great extent, and that the paradigmatic SFG model does permit the generation of syntagmatic, semi-fixed expressions in terms of systemic dependency of a structural or lexical nature (see Tucker (1996) for a detailed analysis of the expression *I haven't the faintest/foggiest/remotest/slightest idea/notion*).

Concept 3.6 Hunston's summary comparison of systemic grammar and phraseology

- a linear view of grammar prioritising lexical dependency can complement a hierarchical view prioritising type of clause combination
- a focus on syntagm, in the sense of sequence, can complement a focus on paradigm, and that the two views can offer each other useful information
- patterns might inform networks, but networks with grammar as a starting point and those with a lexical starting point might not arrive at the same result
- meaning may lie in phrases rather than in systems
- choice may lie between phrases as well as between system alternatives

(Hunston 2006: 76)

3.2.2 SFG and corpus analysis

Just as Tucker (ibid.) has tested the robustness of the system networks on semi-fixed phrases to determine whether they can cope with these, de Beaugrande (1997), taking up a point raised by Halliday, has called for testing on a large corpus to see if the functional descriptions stand up.

Quote 3.8 Corpora and SFG

> The most decisive test for a functional description, as Halliday himself has recently said, is still in front of us: to apply it to a large corpus – to see if its terms and concepts do, in the event, underwrite such wide coverage with a productive convergence. The test will consume years of work by large teams of linguists using sophisticated software. The corpus will be steadily annotated with local descriptions of its grammar and lexicon, gradually converging toward a steadily finer approximation of a global description.
>
> (de Beaugrande 1997: 254)

One major initiative where the SFG approach has been applied to a large corpus of text is the project by the Systemic Meaning Modelling Group at Macquarie University (http://www.ling.mq.edu.au/), one of whose aims is to develop computation tools for text analysis. A suite of tools in their SysAm program can search for various systemic categories, allowing for an increase or decrease in the level of delicacy of the analysis as required by the researcher (Matthiessen 1996) (Table 3.3).

Example 3.2 Modelling material happening clauses using SysConc

Table 3.3 Material happening clauses – pattern of Process + Range: construing motion, derived from SysConc (around 90,000 words of travel text)

Medium/actor	Process	Range/scope
'you'	cross	(main) road, [name] Road, [name] Street
'you'	descend	stairs
'you'	enter	foyer, the Domain
'you'	follow	(main) road, [name] Street, pathway, arrow

(Matthiessen 1996)

As noted in Halliday (2004: 22) the 'grammar is very much harder to get at' in a corpus as a corpus is set up lexically, accessed via the word, thus lending itself much better to lexicographic rather than grammatical analysis. The SysAm tools can therefore be seen as providing a solution to the dilemma posed by Halliday (1992) on the difficulties of querying a corpus for systemic-based grammatical categories.

Quote 3.9 Halliday on the difficulty of retrieving systemic categories from a corpus

> The lexicologist's data are relatively easy to observe [in a corpus]: they are words, or lexical items of some kind, and while their morphological scatter is a nuisance, involving some cumbersome programming and also some awkward decisions, it is not forbiddingly hard to parse them out. The grammarian's data are very much less accessible: I cannot even today ask the system to retrieve for me all clauses of mental process or marked circumstantial theme or high obligation modality.
>
> (Halliday 1992: 64)

Although the SFG and phraseological approaches can be seen, on the whole, as complementary to one another, insofar as they are both looking at the same thing but from different ends of the spectrum, the role of the corpus is somewhat different in each. In the phraseological approach the corpus is used as a reservoir from which to draw language descriptions. In the SFG approach, on the other hand, the corpus tends to be used as a test bed for the SFG system networks and their associated computational tools (see Section 4.5.1 for a discussion on the application of SFG to multimodal text).

The following section examines another approach to corpus linguistics, which also makes use of various software tools specifically designed to accord with the tenets of a particular orientation to language analysis.

3.3 Multidimensional approach

Biber's work has been extremely influential in directing corpus research both in the States and other parts of the world. The theoretical and methodological underpinnings to Biber's MD approach are laid out in his seminal volume *Variation across Speech and Writing* (Biber 1988). More recently, this computational approach to corpus analysis has been extended to the identification of 'lexical bundles' and 'vocabulary-based discourse units' (VBDUs), also discussed in the following subsections.

3.3.1 Multidimensional approach: theory and methodology

Concept 3.7 Multidimensional (MD) approach

Biber (1988) describes a computational, multidimensional statistical analysis for the study of linguistic variation to determine different text typologies. Biber (1988: 73–5) first extracted 67 'linguistic features' in a 1-million-word sample from LOB and the LLC. These 67 features included both syntactic (e.g. tense and aspect markers) as well as lexical features (e.g. specialised verb classes), and also categories such as prepositions and modal verbs, which for some linguists sit uneasily in either category. Biber (2003) has greatly expanded the original inventory from 67 items to 129 linguistic features, with many new lexicogrammatical additions (e.g. *that*-clauses controlled by a noun: *the proposal that he put forward was accepted*).

In the next stage, using factor analysis, Biber identified patterns of co-occurrence of these linguistic features to establish the following five major dimensions across which texts could vary:

1. Informational vs involved production
2. Narrative vs non-narrative discourse
3. Situation-dependent vs elaborated reference
4. Overt vs non-overt expressions of persuasion
5. Non-impersonal vs impersonal style

In this way, Biber established eight text typologies based on the syntactic and lexical features they have in common and then interpreted these co-occurrence patterns to assess their underlying situational, social and cognitive functions. Biber has therefore derived his classification of text typologies from *internal* linguistic criteria, rather than *external* criteria such as communicative function or domain of use. This means that for Biber texts from the same register (taking register to be 'named varieties in a culture, defined in situational terms, like conversation, letters, textbooks, and lectures', Biber et al. 2003: 153) will not belong to the same text type if they do not share the same linguistic features. As an example of the latter point, Biber et al. (2003) mention that their MD analysis of university textbooks and newspaper prose showed these two different registers to be similar according to many of their linguistic characteristics.

One example of the MD analysis applied can be found in Biber et al. (2002), which exemplifies how factor analysis was used to plot the features of various written and academic university registers across the five different

(*continued*)

dimensions. Linguistic features identified for the persuasion dimension are given below (p. 16):

Dimension 4: Overt expression of persuasion

Feature	*Example*
Positive features (overt expression of persuasion)	
Infinitives	hope to go
Prediction modals	will, would, shall
Suasive verbs	command, insist, propose
Conditional subordination	if you want
Necessity modals	must, should, have to
Split auxiliaries	should *really* be
(possibility modals)	can, could, might

This dimension showed there to be a strong polarisation between spoken and written registers, with classroom management and office hours found to be at the top of the overtly persuasive ↔ not overtly persuasive scale, and course packs and textbooks at the bottom end.

Although Kennedy (1998), Lee (2001) and Ghadessy (2003) all acknowledge the importance and usefulness of the MD approach, they have also identified a few problematic aspects in the model. Both Lee (2000), who replicated Biber's study, and Kennedy (2003) share concerns about some of the text type labels on account of their indistinctiveness. For example, Kennedy notes that the 'learned' and 'scientific' types of exposition do not seem to be clearly differentiated and that they may differ only in some cases because of a higher number of active verbs in the 'learned' category.

In spite of these critiques, it cannot be denied that the MD analysis has provided very valuable lexico-grammatical and discourse-based data for the compilation of an empirically based grammar of spoken and written English (cf. Biber et al. 1999). Notably, the MD approach has been applied to analysis of a 60-million-word corpus of professional genres across four disciplines (psychology, social work, industrial chemistry, and construction engineering) in Spanish (Parodi 2007, 2010), and also call centre interactions (Friginal 2009).

3.3.2 Lexical bundles

The original 1988 MD model has been expanded to include 'lexical bundles', which have much in common with Altenberg's study of recurrent word combinations in terms of concept and methodology (see Example 3.1).

Concept 3.8 Lexical bundles

Lexical bundles are another type of phraseological unit, which are identified by purely 'frequency-driven' means. Biber et al. (1999) regard them as recurrent expressions, i.e. 'sequences of word forms that commonly go together in natural discourse' (p. 990). Like Altenberg, Biber states that the frequency cut-off used to identify lexical bundles is somewhat arbitrary, but sets a higher cut-off point for four-word than for five- and six-word bundles, which are much less common. Given the fact that bundles are identified by purely statistical rather than intuitive means, they are more often than not structurally incomplete; for example, *the nature of the* consists of an incomplete noun phrase incorporating the beginning of an embedded *of*-phrase. Biber et al. (1999) also note that shorter bundles are often incorporated into more than one longer lexical bundle, e.g. *I don't think* is also part of *well I don't think* or *I don't think so*. (Altenberg's classification, however, is somewhat different with *I don't think* viewed as a core recurrent sequence with optional extensions, e.g. *well, oh*.)

The following initial taxonomy of lexical bundles is given in Biber (2003: 54). *WH*-initial bundles are common in conversations, while *It*-initial bundles feature frequently in academic prose.

Preposition-initial lexical bundles (e.g. in the form of …)
Other lexical bundles (e.g. the last day of class …)
Noun phrase initial bundles (e.g. those of you who …)
Pronoun-initial lexical bundles (e.g. that's pretty much it …)
WH-initial lexical bundles (e.g. what you're saying is …)
It-initial lexical bundles (e.g. it is possible that …)

As far as other languages are concerned, Cortes (2007) has compared lexical bundles in academic history writing in English and Spanish, Tracy-Ventura et al. (2007) have examined lexical bundles in Spanish speech and writing, and Kim (2009) has carried out a comparison of Korean lexical bundles in conversation and academic text.

Although Sinclair (2002: 353), in a review of the *Longman Grammar of Spoken and Written English*, commends the authors for bringing phraseology to the fore: 'the authors of LGSWE deserve credit for setting up this major signpost, and indicating that a grammar must remain aware of lexis, and that the patterns of lexis cannot be reconciled with those of traditional grammar', Sinclair critiques the concept of lexical bundles on theoretical-descriptive grounds.

Sinclair notes that the concatenation of words [lexical bundles] is shoehorned into existing grammatical categories (a case of data to fit the description rather than letting categories suggest themselves to fit the data), without

any recognition of variability of exponents and discontinuity. Sinclair also queries the arbitrary minimum of three words to a string, which account for 20 per cent of the corpus, pointing out that many existing two-word strings could become three-word strings if the corpus (40 million words) was increased by an order of magnitude.

Both Biber and Sinclair make reference to lexico-grammatical patterns, and on the surface, Sinclair's notion of collocational frameworks seems to overlap with lexical bundles, but in fact, these two aspects are fundamentally different. Whereas lexical bundles can include what Biber et al. (1999) refer to as 'incomplete structural units' (e.g. *result of the*), collocational frameworks are derived from preselected patterns (e.g. *a* + ? + *of*; *an* + ? + *of*; *for* + ? + *of*; *many* + ? + *of*), which are shown by statistical means to have different degrees of productivity with different frameworks attracting different sets of nouns (Renouf and Sinclair 1991; but see Biber 2009 for a study of the productivity of frameworks such as *in the* * *of*). The purely statistical 'frequency-driven' approach is therefore quite distinct from Sinclair's more 'corpus-driven' approach, combining automatic with manual-aided investigations, where meaning has primacy.

However, in spite of Sinclair's misgivings, Biber et al.'s (2004) research has thrown up some illuminating findings on form/function correlations and register variation (see Section 4.3.3 for a functional taxonomy designed to capture the discourse-based nature of lexical bundles). For example, it was found that classroom teaching makes extensive use of noun phrase/prepositional phrase-based referential bundles (e.g. *the nature of the*, *in the absence of*), a particularly surprising finding given that these were found to occur more frequently in classroom teaching than in academic prose, as one would normally expect based on one's 'perceptual salience'. Biber et al. then go on to extrapolate from the findings that these frequent patterns provide evidence that 'lexical bundles are stored as unanalyzed multi-word chunks, rather than as productive grammar constructions' (p. 400) (see Section 3.1.2 on the psycholinguistic salience of formulaic language).

3.3.3 Vocabulary-based discourse units (VBDUs)

Biber's (2004) work on what he terms vocabulary-based discourse units (VBDUs) seeks to bridge the gap between more quantitative corpus-based work where the focus is usually on the surface linguistic analysis of texts and registers, and the more qualitative discourse-based research with its concentration on the internal discourse organisation in a small number of texts.

Concept 3.9 Vocabulary-based discourse units

A computational procedure, namely TextTiling, is used to automatically identify VBDUs. In brief, this is carried out by comparing the words used

in adjacent segments of a text in 50-word batches to see at which point in the text the two adjacent text segments are maximally different in their vocabulary. Where discourse segments are found to be maximally different, such shifts in vocabulary are seen as corresponding to shifts in purpose or topic. At the next stage of analysis, these VBDUs are analysed linguistically using the MD approach outlined previously and cluster analysis is then used to group the VBDUs into groups which are maximally similar.

Ghadessy (2003) has critiqued the 'unit of analysis' in the initial MD methodology on account of the fact that it fails to consider thematic analysis and measurement of the information flow in the text. However, it would seem that this undertaking to combine corpus-linguistic and discourse-analytic research perspectives would serve to address this concern to some extent.

3.4 Sociolinguistic approach of the Nottingham School

One attribute that the approaches discussed above have in common is that they are all sociolinguistically motivated by virtue of the fact that they utilise corpus data. As Stubbs (2001a: 221) notes, 'it [corpus linguistics] is inherently sociolinguistic: the data are attested texts, real acts of communication used in a discourse community'. Sinclair's phraseological approach and Halliday's SFG approach, both deriving from the socio-semantic linguistic theory of Firth but developing in different directions as the above discussion exemplifies, are concerned with how language construes social reality. Biber's multidimensional analysis framework is also socially motivated with its application to different varieties of registers to uncover language use in different social situations. The fourth approach to analysing corpora is that of the 'Nottingham School'. Of all the major approaches to corpus linguistics, this one could be considered the most overtly sociolinguistic on account of its holistic treatment of the 5-million-word Cambridge and Nottingham Corpus of Discourse in English, CANCODE, outlined below.

3.4.1 Contexts and interactional types in a sociolinguistic corpus

The CANCODE corpus of informal spoken English was established in 1996 by Carter and McCarthy (see McCarthy 1998, Chapter 1 for an overview of the corpus). Both the design of this corpus and its subsequent analyses mark a watershed in the era of computerised spoken corpora. Whereas before the 1990s spoken corpora focused on the collection of demographic data, targeting specific populations, and defining the speech categories in terms of 'written-to-be-spoken' or 'rehearsed spoken', for example, Carter and McCarthy's approach

is fundamentally different. Although CANCODE is marked up for demographic data, it has as its main organising principle four broad genre contexts in which spoken language is used (transactional, professional, socialising, intimate), together with a pedagogic sub-corpus (see Section 6.6.3), and focuses on three types of goal-oriented exchanges (provision of information, collaborative tasks and collaborative ideas).

Concept 3.10 Generic organisation and speech genres in CANCODE

The data collected and transcribed for the CANCODE corpus are classified along two main axes according to *context type* and *interaction type*. Context type reflects the interpersonal relationships that hold between speakers, embracing both dyadic and multi-party conversations, and in all cases it is the relationship between speakers, that is, their wish to communicate at this level, which qualifies data for inclusion in the category, and not simply the particular environment in which the audio recording is made. Four broad types are identified along a cline from *transactional*, *professional*, *socialising* to *intimate* (Table 3.4).

Table 3.4 Contexts and interactional types in CANCODE

Context type	**Interaction type**		
	Information provision	*Collaborative task*	*Collaborative idea*
Transactional	Commentary by museum guide	Choosing and buying a television	Chatting with hairdresser
Professional	Oral report at group meeting	Colleagues window dressing	Planning meeting at place of work
Socialising	Telling jokes to friends	Friends cooking together	Reminiscing with friends
Intimate	Partner relating the story of a film seen	Couple decorating a room	Siblings discussing their childhood

(Carter 2004: 149–50)

3.4.2 Sociolinguistic-motivated analyses

Carter and McCarthy's approach to analysis is eclectic, drawing as it does on speech act theory, politeness theory, accommodation theory, praxis theory and conversation analysis (CA), as elaborated on and illustrated in the following discussion.

McCarthy and Carter view language as essentially 'strategic motivated choices', underpinned by speech act and politeness theories. Politeness theory, in Brown and Levinson's model (1987), predicts that speakers will use

more elaborate and more indirect forms of language (for example in making requests) when they want to reduce the level of threat that such actions impose on people's freedom of action (their 'negative face') or their good standing (their 'positive face')'. Evidence in support of 'face' theories is provided by both quantitative and qualitative data from CANCODE.

Quote 3.10 Quantitative data on speech acts from CANCODE

Real data usually show speech acts to be far more indirect and subtle in their unfolding [than invented examples]. Disagreement is a good case. In the CANCODE corpus, there are only eight occasions where someone says *I disagree*, and none where *with you* follows. All eight occasions have some sort of modification which suggests a reluctance on the part of the speaker to utter such a bald statement; these include *I just disagree* (context: semi-formal meeting), *you see now I do disagree, I'm bound to disagree, I'd er, I'd disagree*. Where the verb-form *disagree* occurs, the contexts mostly either 'report' (or predict) disagreement with someone, or disagree with ideas and propositions, rather than people. … It would perhaps be reasonable to assume that other speech acts behave in this way too, unfolding indirectly and in negotiation, with due sensitivity to interlocutors' personal face.

(McCarthy 1998: 19)

Example 3.3 Qualitative data on speech acts from CANCODE

(3.5)

[A young daughter, <S 01>, is being helpful and offering to make everyone toast. Most family members accept two slices. She then addresses her father.]

<S01> Dad, one piece or two?
<S02> **One'll do for me Jen**, if you
<S01> Right, okay.
<S02> Cos I've gotta go in the bath in a minute, love.

… Clearly, speakers wish to avoid the over-blunt refusals … and considerable lexical effort is expended in elaborating the dispreferred response [indicated in bold]. In (3.5), there is a risk that the daughter will interpret her father's choice of only one piece of toast as a snub to her efforts to help in the kitchen. We have what looks like an aborted polite conditional and a reason from the father.

(McCarthy 1998: 56)

Accommodation theory holds that 'speakers will reduce linguistic differences between themselves and other speakers if and when they actively want to communicate more efficiently with them, or to "move closer" in relational terms' (Coupland 2001: 11). Carter and McCarthy map this relational perspective onto Sinclair and Coulthard's (1975) structuralist-functionalist model of the exchange unit of interaction (initiation, response, follow-up), as exemplified below.

Example 3.4 Relational/interactional function of exchange unit (CANCODE)

In the example below, McCarthy notes that the follow-up function very often has a relational/interactional function, 'where social, cultural and affective meanings are encoded in relation to responses, in addition to acknowledging the response and its information, and where key conversational processes such as convergence are effected'.

(3.2)

<S01>	What time is it?	I
<S02>	Twenty to six.	R
<S01>	Is that all?	R/I2
<S02>	Yeah.	R2
<S01>	Oh I thought it was later.	F

(McCarthy 1998: 53)

The analytic approach adopted towards the CANCODE data also draws on praxis theory, which itself implicates the notion of adjacency pairs, turn taking, turn boundaries and sequencing associated with CA. However, as exemplified in the above analyses, Carter and McCarthy extend the boundaries of both speech act theory and CA by interpreting the corpus data both in relation to particular participant motivations for a specific speech event through qualitative analysis, and in relation to socio-discourse norms identified through more quantitative analyses.

Concept 3.11 Definition of praxis theory

In praxis theory, analysts … hold that the outcomes of talk are largely unforeseeable, that talk or conversation develops its own momentum, and that meanings are therefore *contingent* (they depend on other meanings around them) and *emergent* (they surface progressively and incrementally from the flow of talk). Agency tends to be construed as shared between participants, so meanings and talk itself are said to be *co-constructed*, or else, more radically, agency is attributed to the process of social interaction itself.

(Coupland 2001: 12)

The main thread running throughout Carter and McCarthy's corpus-based research is that meanings are negotiated face to face and emerge from the unfolding discourse. And meanings are just as much interpersonally motivated as they are task-oriented, co-constructed within a Vygotskyan framework where 'mind exists in social space'. Various resources used for the social construction of meaning and interpersonal convergence have been shown to include hyperbole (McCarthy and Carter 2004), repetition and relexicalisation (McCarthy 1998), and creative use of language (Carter 2004; Carter and McCarthy 2004).

Based on recurrent patternings in CANCODE, Carter and McCarthy (1995) and McCarthy and Carter (2002) also posit the notion of an *interpersonal grammar* of spoken English, exemplifying how certain grammatical features, e.g. tags and amplificatory noun phrases occupying the tail slot of a sentence: It's very nice *that road up through Skipton to the Dales*, signal relationships between participants and their stance or attitudinal positioning towards the emergent discourse. Section 2.2.1 has already noted the putative gulf between the notion of corpus as *product* and corpus as *process*. The idea that grammar can be interpersonally motivated seeks to bring these two aspects more into alignment.

Quote 3.11 Text grammar vs discourse grammar

> In the Hallidayan paradigm, grammatical choices reflect concerns such as across-sentence cohesion, the creation of texture (the feeling that the text is a coherent whole), and the ways speakers and writers position themselves in texts by choices within grammatical systems, such as modality, transitivity, tense, voice, and so on. … Halliday's grammar is essentially a text-grammar, that is to say, the choices are examined in relation to how the finished product, the text, comes to be as a result of choices made from predetermined systems, whereas discourse grammars are more process-oriented and are interested in any individual interactional factor that may influence moment-by-moment choices in context.
>
> (Hughes and McCarthy 1998: 264)

Although the multifaceted sociolinguistic approach adopted by Carter and McCarthy to spoken corpus data is a powerful and dynamic one, two issues arise: interpretivity and universality. Information on speaker relationships has been considered to aid the analyses in CANCODE, but it still remains that the corpus analyst's role is an interpretative one (see Section 1.3.2) and some pragmatic features of discourse may remain rather elusive. In this respect, as far as hyperbole is concerned, McCarthy and Carter (2004: 178) note that although 'There is limited evidence that can be drawn concerning speakers' "pragmatic sincerity" from a corpus', they maintain that it is the *cumulative* effects of turn boundaries, extreme formulations, etc. and the listener's contributions to the

interactive emergent discourse that can 'all point to speakers' intentions to highlight contrasts between expectation and reality'. The suggestion, here, that process features can be extrapolated from various configurations of the corpus data is an intriguing one which merits further investigation in corpus exploration of a qualitative nature.

Another area for further research concerns the extent to which the interpersonal features identified in the CANCODE corpus can be considered universal, and not just an artefact of that particular corpus. Indeed, McCarthy (1998: 21) calls for duplication 'for any dialect or sociolect (including "global/international" English)' in order to decide what could be classified as 'standard features' of spoken language (see Section 6.1.1 for discussion of the role of corpora in defining English as a lingua franca).

The fine-grained sociolinguistic analyses and delicate categorisations of Carter and McCarthy's approach to spoken data are therefore quite different from the broad-brush quantitative approach adopted by Biber et al. (1999), in which conversational data constitute one of the superordinate categories for register analysis, the others being fiction, newspaper language and academic prose.

In the following chapter, I discuss key corpus-based discourse-oriented studies which draw on the theoretical/methodological stances of the approaches discussed above.

Further reading

Biber, D. (2006) *University Language. A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins. For readers wishing to have an overview of Biber's work, this is a good starting point. The volume is highly accessible with clear examples of corpus data.

Gilquin, G. (2010) *Corpus, Cognition and Causative Constructions*. Amsterdam: John Benjamins. This work combines corpus linguistics with cognitive linguistics to identify the syntactic, semantic, lexical and stylistic features that are distinctive for individual causative constructions.

Hunston, S. (2010) *Corpus Approaches to Evaluation*. London: Routledge. This volume shows the importance of phraseology for the study of evaluative language.

Stubbs, M. (1996) *Text and Corpus Analysis*. London: Blackwell. Chapter 2 provides an excellent account of British traditions in text analysis (Firth, Halliday and Sinclair).

Thompson, G. and Hunston, S. (eds) (2006) *System and Corpus. Exploring Connections*. London: Equinox. This edited collection presents an eclectic mix of corpus-based studies in the SFG tradition.

Part II
The Nexus of Corpus Linguistics, Textlinguistics and Sociolinguistics

4
How is Corpus Linguistics Related to Discourse Analysis?

This chapter will discuss:

- Different conceptual orientations towards corpus linguistics
- Differences between corpus analysis and discourse analysis
- Different approaches to analysing written, spoken and multimodal corpora
- Challenges posed in analysing 'new technologies' corpora

4.1 Is corpus linguistics a theory, a methodology or an approach?

Around two decades ago, corpus linguistics was considered a data analytical methodology since it could speed up and systematise enquiries into lexis, grammar or lexico-grammar. Citing Leech (1992), Tognini Bonelli (2001: 48) states that '… the corpus advanced the methodology but did not change the categorical map drawn by linguistic theory'. Tognini Bonelli then goes on to suggest that some linguists came to regard corpus linguistics as going beyond a methodology for enhancing linguistic analysis to take on a more evolutionary role, bringing together linguistic theory and data, making possible what Halliday termed 'probabilistic grammars' (1991, 1992, 1993).

Quote 4.1 Corpus as theory

> … what had started out as a methodological enhancement … has turned out to be a theoretical and qualitative revolution in that it has offered insights into the language that have shaken the underlying assumptions behind many well established theoretical positions in the field. … It is strange to imagine that just more data and better counting can trigger philosophical repositionings, but that is what these writers felt, and that indeed is what has happened.
>
> (Tognini Bonelli 2001: 48)

Tognini Bonelli thus sees corpus linguistics as a catalyst in redefining aspects of linguistic theory through a 'corpus-driven' approach, in which the data are approached without any preconceived notions in relation as to how they should be analysed. As discussed in Section 3.1, corpus linguists advocating a 'corpus-driven' approach oppose any a priori mark-up to aid analysis as this would obscure, they argue, any potential new insights into language. Accordingly, Tognini Bonelli (2002: 75) calls for a change in the *unit of currency*, i.e. the unit of linguistic investigation, echoing Sinclair's (1985) view that there needs to be an overhaul of the present descriptive systems for deriving a new theory of language. In fact, it is surprising how long it took for this view to be commonly recognised, and for corpus work to be seen as connecting with theoretical issues, an observation made by Hall and Beggs (1998). Citing Tyler (1987: xi), Hall and Beggs note that this 'radical change of focus' has overturned the idea of language simply viewed as a rule-governed system: 'The easy assumption of the old order of discourse – of wholeness, consensus, clarity, closure, telos, and even order itself – now seem awkward, unfamiliar and almost embarrassing' (p. 31). However, Hall and Beggs also draw attention to the reluctance on the part of the linguistic community to accept this new order, quoting Strecker (1985: 24): 'Facing chaos, most people seem to stick to anything that promises order and plain answers. And above all, they stop asking questions that might push them back into confusion.'

However, most linguists take a less extreme or slightly different position from that of Sinclair, with Butler (2003b: 132) viewing corpus data not merely as a repository to provide examples to justify theoretical claims, but rather as a means for rigorously testing those claims 'leading to the modification and if necessary the abandonment of particular proposals'. Butler's view is not quite as radical as that of Tognini Bonelli, lying somewhere between the 'corpus-based' and 'corpus-driven' approach. Likewise, Aarts (2002b) views corpus linguistics as a methodology for validating existing descriptions of language on which to make changes in the description and annotation scheme where the corpus data do not fit these.

Neither do McEnery et al. (2006) see such a gulf existing between the two approaches, and indeed, argue that, in essence, there is no substantial difference between the two. They note that traditional categories such as nouns, verbs, prepositions, subjects, objects, clauses and passives are found in those studies identified as 'corpus-driven'. Such categories have therefore not been completely abandoned, indicating, as McEnery et al. state, that in the 'corpus-driven' approach linguists are still applying intuitions based on traditional grammatical terms, in line with Popper's (1963) view that there exists no completely theory-free observation. While McEnery et al. (ibid.: 6) consider corpus linguistics as 'a new philosophical approach to linguistic enquiry' with its own theoretical status, they do not view it as a discipline in its own right with its own theory. Teubert (2005: 2) describes the field as 'a theoretical approach to the study of language'.

Quote 4.2 McEnery, Xiao and Tono on status of corpus linguistics

> As corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory itself. The qualitative methodology used in social sciences also has a theoretical basis and a set of rules relating to, for example, how to conduct an interview, or how to design a questionnaire, yet it is still labelled as a methodology upon which theories may be built. The same is true of corpus linguistics.
>
> (McEnery et al. 2006: 7–8)

In sum, there are many competing viewpoints as to whether corpus linguistics should be considered a methodology, theory or approach. It is probably best regarded, in essence, as a methodology along the continuum (rather than divide) of the corpus-driven vs corpus-based approaches. Although the research results are being increasingly interpreted with reference to other linguistic (e.g. systemic-functional linguistics, see Section 3.2) or cognitive theories such as those embraced by usage-based models of language (see Section 3.1.3), this does not make corpus linguistics a theory in itself. For this reason it may be more appropriate to refer to this field as 'corpus-based linguistics', as Lee (2008) does, to clarify its status.

4.2 Corpus analysis vs discourse analysis

We have seen in Section 4.1 that corpus linguistics is a somewhat slippery term to define and that there has been much debate over corpus-driven vs corpus-based investigations (see Section 3.1). Likewise, discourse analysis is a field of enquiry whose essential nature, like that of corpus linguistics, has also come under scrutiny.

Quote 4.3 Summary of approaches to discourse analysis

> Discourse analysis covers a vast range of areas and is also one of the least clearly defined. (Aijmer and Stenström 2004a; Stubbs 1983). Blommaert (2005: 2) notes that traditionally discourse has been treated in linguistic terms as 'language-in-use', informing areas such as pragmatics and speech act theory. However, for Blommaert discourse has a wider interpretation as 'language-in-action', i.e. 'meaningful symbolic behaviour'. Jucker
>
> (*continued*)

et al. (2009a: 5) define this wider use of the term discourse as 'the totality of linguistic practices that pertain to a particular domain or that create a particular object'. A useful distinction is made by Gee (2001), who defines the 'language-in-use' aspect as discourse (with a little 'd') and the more 'language-in-action' orientation as Discourse (with a capital D), involving not only linguistic practices but other semiotic elements. Discourses are created through recognition work of 'ways with words, actions, beliefs, emotions, values, interactions, people, objects, tools and technologies' (Gee ibid.: 20) that constitute a way of being a member of a particular discourse community.

(L. Flowerdew 2011b)

Although discourse analysis and corpus linguistics both make use of naturally occurring, attested data, they have intrinsically ontological and epistemological differences, as noted by Virtanen (2009). Doing corpus analysis is not the same as doing discourse analysis (DA). Leech (2000: 678–80, cited in McEnery et al. 2006) observes that there is a 'cultural divide' between the two: 'while DA emphasizes the integrity of the text, corpus linguistics tends to use representative samples; while DA is primarily qualitative, corpus linguistics is essentially quantitative; while DA focuses on the contents expressed by language, corpus linguistics is interested in language *per se*' (p. 111). Tognini Bonelli (2004) also notes that corpus linguistics is not the same as doing text analysis (Concept 4.1).

Concept 4.1 Corpus analysis vs text analysis

A text	*A corpus*
Read whole	Read fragmented
Read horizontally	Read vertically
Read for content	Read for formal patterning
Read as a unique event	Read for repeated events
Read as an individual act of will	Read as a sample of social practice
Coherent communicative event	Not a coherent communicative event

(Tognini Bonelli 2004: 18)

The main epistemological differences between the two fields lie in the fact that corpus analyses, by virtue of their methodological status, treat the text as a product rather than as an unfolding discourse as process and social action: '… the computer can only cope with the material products of what people do when they use language. It can only analyse the textual traces of the processes

whereby meaning is achieved' (Widdowson 2000: 4). See Section 3.4.2 where it is argued that the notion of an interpersonal grammar does attend to process to a certain extent.

As far back as 1998, I drew attention to the potential of corpus linguistics for 'doing' discourse analysis. McEnery et al. (2006: 111) state that the aforementioned cultural divide 'is now diminishing', and Partington (2004b) proposes that corpus and discourse methods are complementary. This following section seeks to examine to what extent corpus and discourse approaches have now established a common meeting point, given their inherent differences in epistemologies and methodologies. Studies of corpus-based discourse analyses will be discussed from the following four perspectives, which subsume different, yet overlapping, theoretical underpinnings.

Written corpora

- *Genre-based*: an approach which focuses on language choices, meanings and patterns in texts including those based on the Swalesian (2004) notion of genre and also the New Rhetoric approach.
- *Problem–solution based*: an approach which examines aspects of the four-part situation–problem–solution–evaluation pattern (Hoey 2001).
- *Linguistic devices with discourse functions*: e.g. lexical bundles, metadiscourse expressions and metadiscourse nouns.
- *Critical discourse-based*: an approach which brings an attitude of criticality, such as critical discourse analysis (CDA), but also draws on other methods, e.g. SFL.

Spoken corpora

- *Prosodic*: corpus investigations which focus on the relationship between prosodic information (e.g. tone choices, stress, pauses) and discourse features.
- *Rhetorical*: corpus investigations which focus on pragmatic devices such as hedges.

Multimodal corpora

These corpus studies involve text which is accompanied by sound and video files. Three different approaches to compiling and analysing multimodal corpora can be discerned in the literature:

- SFL-inspired
- Functional
- Situated discourse

Hybrid corpora

These corpora are associated with the 'new technologies' involved in computer-mediated communication, such as blogs.

Corpus-based discourse analyses can be viewed not only from the four different modes above together with their attendant discourse areas, e.g. genre-based or SFL-based approaches, but can also be seen in terms of subject areas, i.e. workplace discourse, media discourse, academic discourse, etc., and as a reflection of certain ideological positionings, i.e. discourses of racism, and discourse of power. Moreover, it should also be noted that many of the discourse-based studies cited below implicitly subscribe to the 'corpus-driven' approach with their focus on the phraseological nature of language in which the lexical item has primacy (see Section 3.1).

4.3 Discourse analysis: written

Corpus studies discussed in this section can be classified according to whether they are primarily genre-based, problem–solution based, examine linguistic devices from the perspective of discourse functions, or are CDA-based. While many have a mainly text-based, i.e. language-in-use focus, at the same time, they also address the interpersonal nature of language such that the analyses are reader- and/or writer-oriented and take situational and contextual features into account.

4.3.1 Genre-based approaches

Three different theoretical positions on genre can be identified in the literature, namely: (1) the Swalesian tradition of genre associated with English for Specific Purposes (ESP) texts, (2) North American New Rhetoric studies, and (3) Australian systemic-functional linguistics (see Hyon 1996 for an overview of these three approaches). However, this chapter will concentrate on the first two approaches in written text as these are the genre traditions to which corpus linguistic techniques have been applied (see Handford 2010a for an overview of genre-based corpus studies across professional, academic and non-institutional discourse).

Swalesian tradition of genre

Corpus studies motivated by the Swalesian notion of genre as a goal-driven communicative event associated with particular discourse communities are discussed below. J. Flowerdew and Forest (2009) apply Swales' (1990: 141) CARS ('Create a Research Space') model, originally posited for academic research article introductions, to PhD literature reviews, investigating the patterning of the keyword 'research' across different moves and steps. For example, it was found for the 'gap-indicating' move that 'research' was involved in two canonical patterns: *There has been little research/little research has been done*; *Further research is needed/called for* (p. 27).

Another genre-motivated study is that by Bhatia et al. (2004) on law cases. A classification into genre moves of the verbs *submit*, *dismiss*, *reject* and *grant*, commonly found in law cases, shows that they clearly have a preference for different move structures (Table 4.1).

Example 4.1

Table 4.1 Verbs occurring in genre moves in law cases

	Frequency			
Genre move	**Submit**	**Dismiss**	**Reject**	**Grant**
1 Facts/stating history of the case	75	**47**	12	**82**
2 Presenting argument	**263**	6	9	51
3 Deriving *ratio decidendi*	5	16	**44**	**80**
4 Pronouncing judgment	3	**42**	9	16
Total	**346**	**111**	**74**	**229**

(Bhatia et al. 2004: 214)

However, Bhatia et al.'s (2004) study of genre moves in law cases reveals the limitations of a purely corpus-analytic approach: in order to make a pragmatic distinction between seemingly synonymous verbs, such as *dismiss* and *reject* in law cases, Bhatia et al. (ibid.: 213) state that it would be necessary to 'look for evidence from institutional practices' as corpora cannot (usually) provide such information.

While most genre-based corpus studies commence from a lexico-grammatical, bottom-up perspective, Durrant and Mathews-Aydınlı (2011) argue for a 'function-first', i.e. top-down, approach, maintaining that the communicative context needs to be integrated into the analysis from the outset with identification of formulaic expressions grounded in the pre-identified functions. Kanoksilapatham's (2005, 2007) research takes a rhetorical top-down perspective at the outset. In her study of biochemistry research articles Kanoksilapatham first develops an analytical genre-based framework through the identification of rhetorical move types and then uses Biber's multidimensional (MD) analysis to determine the linguistic characteristic of each move, as illustrated below (see Section 3.3 for an overview of Biber's MD analysis).

Example 4.2 Kanoksilapatham's analytic procedure for investigating stance

Kanoksilapatham first devised a model of move structure in biochemistry research articles. Four types of moves were identified in the Results section, for example:

- Move 8: Restating methodological issues
- Move 9: Justifying methodological issues

(*continued*)

- Move 10: Announcing results
- Move 11: Commenting results

Each move was then broken down into steps. For example, the steps identified for Move 11: Commenting results are as follows:

- Move 11: Commenting results
  Step 1: Explaining results
  Step 2: Generalising/interpreting results
  Step 3: Evaluating results
  Step 4: Stating limitations
  Step 5: Summarising

Based on Biber's (1988) procedure for identifying functionally motivated dimensions for text analysis, Kanoksilapatham then came up with seven dimensions for biochemistry research articles, one of which denotes 'evaluative stance'. The co-occurring linguistic features for this communicative dimension would be the extraposed '*it*' construction providing a means for authors to express their comments or attitudes in a somewhat inexplicit way. Predicative adjectives also provide a means for authors to express their stance, as do *that* complement clauses and *to* complement clauses controlled by adjectives.

Kanoksilapatham then plotted the move structures for each of the sections in the report, according to their linguistic characteristics, along the 'evaluative stance' continuum. Not surprisingly, the two moves with the lowest dimension score for evaluative stance were those describing experimental procedures and materials, while the step in the move structure with the highest score was that stating the limitations of the study, e.g.

> In the absence of an atomic structure, **it** is not *possible* to determine which residues are solvent exposed and thus likely to make physical contact with the microtubule and which ones contribute to the domain's structural organization.
>
> (Adapted from Kanoksilapatham 2007: 76–81)

Corpus-based studies, such as Kanoksilapatham's, are of value not only for identifying those moves which seem to be typical of a genre, but also those that are somewhat unconventional. She identified additional moves in the Results section where the data are not only reported but also commented on (e.g. *We presume that … is representative of … because*). This deviates from the style

prescribed in publication manuals which stipulate that all subjective evaluation and comments should be left to the Discussion section. The rationale for this arrangement is provided by Swales who notes that the introduction of evaluations and justifications at an earlier stage than expected can in appropriate circumstances 'work to impress and reassure the reader that the paper is worth pursuing further' (Swales 2004: 232).

New Rhetoric approach

The New Rhetoric approach to genre differs from that associated with the Swalesian approach in that it is not concerned with linguistic features per se. However, in common with this approach it also places a great deal of emphasis on the social purposes that genres fulfil in certain situational contexts, viewing genres as dynamic texts which are shaped and influenced by other related texts and utterances of the sociocultural context (see Bazerman 1988, 1994; Freedman and Medway 1994), a view in line with Bakhtin's (1986) notion of intertextuality (see Section 6.6.2 for studies which account for aspects of learner language through the influence of related texts). Scholars working in the New Rhetoric tradition have also tended to use ethnographic (i.e. participant observation and interviews) rather than linguistic or rhetorical methods, such as the move structures, for analysing texts, or combined them both, as have Hyland (1998) and Tse and Hyland (2006). The New Rhetoric approach can also be seen as drawing on activity theory to explore how knowledge is constructed with reference to the larger activity systems and institutional cultures in which it is situated (cf. Engeström and Middleton 1996).

Although Hyland does not refer specifically to the New Rhetoric approach, his research does apply ethnographic methods for interpretation of corpus data. For example, Hyland (1998) consulted specialist informants for his study on the use of hedging devices in a corpus of 80 research articles in cell and molecular biology. To ensure a faithful interpretation of hedging devices, Hyland involved in his analysis four native speaker biologists, who were all experienced researchers and writers in the field, by asking them to voice their reactions to underlined features in the text and also by having them participate in small focus group discussions to elucidate why the 'expert' writing under investigation was appropriately phrased for the readers. In another study on the use of metadiscourse in a corpus of 84 academic book reviews from three contrasting disciplines, Tse and Hyland (2006) interviewed journal editors and reviewers to assist in defining and verifying how experts in this field shaped their pragmatic strategies to adhere to the formal constraints of this genre.

Widdowson (2000: 4) has remarked that corpus-based methods focus on the text as product and 'cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted'. Neither can the computer 'produce ethnographic descriptions of language use' (ibid.: 4). Lee (2008: 95)

also echoes this view, commenting that the lack of immediacy of the discourse vis-à-vis the analyst 'may be a hindrance for the types of discourse analysis that rely on intimate knowledge of the data, participants and context'. However, the more ethnographic approach to analysing corpus data adopted by Hyland serves to offset this criticism to some extent and enables the text to be 'read as a sample of social practice' (see Concept 4.1). Incorporating an ethnographic dimension to studies of an intercultural nature, most of which seek to explain differences between research articles in English and Spanish, would serve to strengthen culture-bound explanations.

4.3.2 Problem–solution based approach

The problem–solution (P–S) model of discourse (cf. Hoey 2001) has also been applied to corpus investigations. L. Flowerdew (2003, 2008a) used the Appraisal system from systemic-functional linguistics (Martin and White 2005) for classifying keywords for the P–S pattern in technical reports into different types of evaluative lexis, followed by microanalyses of the semantic relation of cause and effect. Keywords, i.e. words which are found to occur with greater frequency when compared with a reference corpus, have been the focus of several studies for examining corpora at a textlinguistic level; see Bondi and Scott 2010 for further studies). Ali Mohamed (2007) investigated the problem element in another text type, Malaysian and British journalistic business texts, also applying Martin's Appraisal system for categorising interpersonal and evaluative meanings. A key feature of her study is the use of the *WMatrix* corpus tool (Rayson 2008) to identify different semantic fields characteristic of the problem element.

While the two previous studies are text-based in orientation, one corpus study investigating the P–S pattern in terms of a more reader- and writer-oriented perspective accompanying the textual analysis is that by Alonso Belmonte (2009) of two corpora, newspaper editorials and op-eds (236 of each). An interactional analysis of different communicative acts (e.g. justification, exemplification) associated with different elements of the pattern was complemented by an illocutionary analysis with the corpus coded for speech acts such as assertions, shared-knowledge assertions, etc., indicating how writers conduct interaction with their readers.

4.3.3 Linguistic devices with discourse functions

Section 1.2.3 illustrated how collocations and colligations in Hoey's theory of lexical priming can operate at the textlinguistic level. Three other types of discourse-based devices are discussed in this section, namely lexical bundles as they are realised functionally, metadiscourse and metadiscoursal nouns. A key feature of the corpus studies reviewed below is that they are often contrastive in nature, highlighting variation across different university disciplines, genres, and spoken and written registers.

Lexical bundles

Section 3.3.2 outlined Biber's structural classification of lexical bundles. In this section, they are discussed from a discourse-based functional perspective. Biber et al. (2004) outline three main functional categories: discourse organisers, referential expressions and stance expressions, the latter consisting of epistemic stance bundles which comment on the knowledge status of the information (e.g. *I don't know if*) and attitudinal/modality stance bundles which express speaker attitude towards certain actions (e.g. *I want you to*).

Hyland's (2008) categorisation of lexical bundles is similar to that of Biber's, but differs in that like his classification of metadiscourse (see following section), they are organised around categories which reflect either the writer or reader involvement in the text.

Concept 4.2 Hyland's functional classification of lexical bundles

Research-oriented – help writers to structure their activities and experiences of the real world, e.g.:
Procedure (*the use of the, the operation of the*)
Quantification (*the magnitude of the, the surface of the*)

Text-oriented – concerned with the organisation of the text and its meaning as a message or argument, e.g.:
Structuring signals – text-reflexive markers which organise stretches of discourse (*in the present study, in the next section*)
Framing signals – situate arguments by specifying limiting conditions (*in the case of, with respect to the*)

Participant-oriented – these are focused on the writer or reader of the text, e.g.:
Engagement features – address readers directly (*it should be noted that, it can be seen*)

(Adapted from Hyland 2008: 13–14)

Both Biber's and Hyland's corpus research on lexical bundles have provided insights into the functional differences between spoken and written registers (Biber 2006; Cortes 2004) and disciplinary variation in the academy (Hyland 2008). Of note is that research on lexical bundles commences from a bottom-up perspective, in contrast to the top-down studies by Durrant and Mathews-Aydınlı (2011) and Kanoksilapatham (2007) discussed in the previous section.

Metadiscourse

The most extensive work in the area of metadiscourse has been carried out by Hyland (2004, 2005).

Quote 4.4 Hyland on metadiscourse

> Metadiscourse is self-reflective linguistic expressions referring to the evolving text, to the writer, and to the imagined readers of that text. It is based on a view of writing as a social engagement and, in academic contexts, reveals the ways writers project themselves into their discourse to signal their attitudes and commitments.
>
> (Hyland 2004: 133)

Of note is that the above definition encapsulates and synthesises the three broad approaches to writing (text-oriented; writer-oriented; reader-oriented) outlined in Hyland (2002). As Hyland and Tse (2004: 161) point out, this model of metadiscourse re-evaluates previous work in the field and, as such 'rejects the strict duality of textual and interpersonal functions found in much previous work'. They suggest that all metadiscourse can be viewed as interpersonal, as it encompasses the reader's knowledge, textual experiences and processing needs. For this reason, Hyland (2005) proposes two main categories for classification: *interactive* resources (e.g. transitions and frame markers) to help guide the reader through the text, and *interactional* resources (e.g. hedges and boosters) to involve the reader in the argument (Concept 4.3).

Concept 4.3 Hyland's interactional level of metadiscourse

Interactional	**Involve the reader in the text**	**Resources**
Hedges	Withhold commitment and open dialogue	Might; perhaps; possible; about
Boosters	Emphasise certainty or close dialogue	In fact; definitely; it is clear that
Attitude markers	Express writer's attitude to proposition	Unfortunately; I agree; surprisingly
Self-mentions	Explicit reference to author(s)	I; we; my; me; our
Engagement markers	Explicitly build relationship with reader	Consider; note; you can see that

(Hyland 2005: 49)

Hyland (2005) and Hyland and Tse (2004) looked at variation in metadiscoursal choices across different disciplines in the area of postgraduate dissertations, research articles and academic textbooks. The analyses reveal the extent to which the socio-rhetorical context of writing affects choices at the metadiscoursal level.

Example 4.3 Disciplinary variation in metadiscourse

A comparison of dissertations from six disciplines in the natural and social sciences showed that the more discursive 'soft' fields such as applied linguistics and public administration employed more metadiscourse overall than fields such as computer science and electronic engineering, and accounted for two-thirds of the interactional features (e.g. hedges, boosters, attitude markers, engagement markers, self-mentions). Hyland accounts for this finding in the humanities as follows:

> Dealing with human subjects and data is altogether more uncertain and writers are unable to draw to the same extent on empirical demonstration or trusted quantitative methods. Consequently persuasion lies far more in the efficacy of argument and the role of language to build a relationship with readers, positioning them, persuading them, and including them in the argument.
>
> (Hyland 2005: 58)

Hyland's category of self-mentions (see Concept 4.3) has been investigated in other corpus-based studies of academic discourse (e.g. Harwood 2005a, b). What is noteworthy about Harwood's research on the use of the personal pronouns *I* and *we* is that he appears to be taking a more critical discourse-analytic approach to the data by looking at how personal pronouns combine with co-textual features to assert the authority of the author and act as a means of self-promotion.

Metadiscourse nouns

One area that has received a lot of attention is how certain nouns function at a discourse level. Drawing on Biber et al.'s distinction between epistemic and attitudinal markers, Charles (2003) compares the use of epistemic nouns, e.g. *assumption*, and stance nouns, e.g. *problem*, in postgraduate theses from the fields of politics and materials science.

Concept 4.4 Biber's categorisation of epistemic and stance markers

Epistemic stance:
indicates how certain the speaker or writer is, or where the information comes from (e.g. *assumption, impression, possibility*)

(*continued*)

Attitudinal stance:
indicates feelings or judgements about what is said or written (e.g. *problem, failure, threat*)

(Adapted from Biber 2006: 93)

In Charles' study metalinguistic nouns were found to function retrospectively, thus having an interpersonal function as they indicate to the reader how the proposition is to be interpreted. Like Hyland, Charles (2006) also interprets her findings with reference to the different epistemologies and ideology of the discipline.

Example 4.4 Disciplinary differences in stance nouns

In Charles' two corpora, twice as many metalinguistic nouns of the epistemic kind were found in the politics corpus compared with the materials science one. Where nouns with a discourse-organising role were found in the materials corpus, these were non-metalinguistic (e.g. *effect, result, procedure*).

> The following passages are recorded in the Sub-Committee's Report: [quotation omitted]. *This discussion* between state representatives at the UNHCR illustrates the significant issues in the debate over gender-related persecution. (Politics)

> Specimens were cut, ground and polished, as for the optical examination. *This procedure* was followed in the first instance by electro-polishing … (Materials)

As did Hyland, Charles also interpreted her findings with reference to the socio-rhetorical context of writing:

> We may account for this [the fact that Politics uses twice as many metalinguistic nouns] by referring to different disciplinary differences in the construction of knowledge. First, politics research mostly draws upon resources that are language-based: both written and spoken records. It is also primarily concerned with constructing accounts and interpretations of events that are predominantly in written form. Thus the activity of the discipline is inherently text-based. This leads to a higher frequency of metalinguistic nouns.
>
> (Charles 2003: 321)

Interestingly, the frequency of attitudinal stance nouns was found to be quite low in both corpora, and when occurring in the Politics corpus usually expressed a negative stance toward political situations or events (e.g. *dilemma, restriction, failure*).

Another corpus-based study of nouns with a discourse function is that by J. Flowerdew (2003a, b), who uses the term 'signalling noun' to denote abstract nouns (similar to those discussed in Schmitt 2000), the meaning of which is only fully realised in context.

Concept 4.5 Discourse function of signalling nouns

'Signalling nouns' (e.g. *way, effect, case*) can be realised both within and outside the clause within which they occur, operating anaphorically (i.e. retrospectively) and cataphorically, although the anaphoric function is more common. Differences in the functions of these nouns in academic speech and writing show they are often realised exophorically in spoken discourse:

meaning realised within the clause
The quickest ***way*** of doing this in the case described above is to set up a group of enucleated Amoebae, perhaps fifty in all (the experimental group).

meaning realised across clauses, anaphoric
In the case illustrated in Figure 3.1 the secretion is released from the free surface of the cells. Mucus is secreted this **way**, as is sweat from the sweat glands in the skin.

(J. Flowerdew 2003a: 38)

exophoric: there is nothing earlier or later in the text to realise a specific meaning of "way"
/now this is paramecium stained in a different ***way***/not to show what's inside in cytomplasm/but to show the cilia on the pellicle/

(J. Flowerdew 2003b: 333)

Corpus linguistic techniques have thus proved of great value in shedding light on how various language choices and patterns operate at a textlinguistic level, either at a level above the clause or sentence, or within the framework of discourse models, very often using a multi-pronged approach, e.g. combining Biber's MDA with Swalesian genre move structures. Moreover, the focus on lexis in the above corpus-based studies, especially textlinguistic nouns of various kinds, shows that lexis is no longer 'the poor relation' of grammar (see Schmitt 2010 for a survey of corpora used as a tool in researching vocabulary).

4.3.4 Critical discourse-based approach

In order to situate corpus linguistics in relation to CDA, it is first of all necessary to outline the basic theoretical underpinnings of CDA. By so doing, the insights that corpus linguistic techniques bring to bear on CDA can be better appreciated.

Quote 4.5 Summary of the two main approaches to CDA

> In CDA the focus is on 'discourses', rather than discourse per se, referring to a broad range of linguistic and nonlinguistic social practices and ideological assumptions co-constructing, for example, 'discourses of power' or 'discourses of racism', in other words, Discourse with a capital 'D'.
>
> (Gee 2001)
>
> The techniques of CDA are multi-fold and vary. Text-analytic techniques draw on SFL, pragmatics and speech act analysis, and are integrated with concepts from contemporary social and cultural theory. Thus, CDA is not a method, as such, in itself but rather 'an academic movement', drawing on a kaleidoscope of methods increasingly those associated with corpus linguistics.
>
> (Baker et al. 2008)
>
> Two main approaches to CDA have developed since the 1960s. In the approach associated with Fairclough (2000, 2003) the analytical framework centres on a discursive event, an instance of language use, analysed not only as text, but also as discursive and social practice. The discourse-historical approach associated with the Viennese school (Wodak & Meyer 2009a) takes a more interdisciplinary, sociolinguistic perspective to the data in which ethnography can also be a part of the analytical procedures.
>
> (L. Flowerdew 2011b)

In fact, corpus-based CDA, which attempts to link recurring patterns in text with sociolinguistic features from the original contextual environment and vice versa, is a relatively new field (see Hunston 2002, and Mautner 2009b for a review of key studies). The pioneering work of Stubbs (1996, 2001a) and Hardt-Mautner (1995) has given rise to the newly emerging interdisciplinary field of corpus-assisted discourse studies (CADS), an approach underpinned by Fairclough's concept of CDA.

Corpus-assisted discourse studies (CADS)

Several of these studies examine the pervasive phenomenon of evaluation (cf. Hunston and Thompson 2000), applying Martin and White's (2005) Appraisal

System in Systemic Functional Grammar (see Bednarek 2006 for an in-depth study of evaluation in media text). Coffin and O'Halloran (2005) also make use of the Appraisal system for classifying evaluative lexis, first conducting an individual text analysis and following this up with a large-scale computerised analysis.

Example 4.5 Appraisal system applied to CDA

Coffin and O'Halloran (2005) first carry out a detailed qualitative analysis of a report from *The Sun* using the Appraisal system, specifically judgement, graduation and affect in their classification. For example, in the sentence below, bold indicates graduation, underlining judgement and italics affect:

> **Two million jobs** will be lost if Tony Blair signs the EU treaty (negative indirect judgement of Blair), it was *feared* last night.

They then used a 45-million word newspaper corpus, made up of *The Sun* and its Sunday version, *The News of the World*, to check any potential overinterpretation of their Appraisal analysis. Their concordances of *United States of Europe* reveal many of the local lexico-grammatical environments to indicate a negative evaluation, e.g. … *leader's bleak plan for a United States of Europe came as a hammer blow to* … (see Example 1.5).

Because of the negative prosody of the *United States of Europe*, Coffin and O'Halloran argue that *Sun* readers will be potentially predisposed to evaluate related expressions negatively even when they occur in a seemingly neutral statement, as in the case of the last sentence (*Mr Blair will be expected to sign up to the constitution blueprint by the end of June*) in their text chosen for qualitative analysis (see also O'Halloran 2009).

(Adapted from Coffin and O'Halloran 2005: 149–57)

Much work in this area has also been carried out under the auspices of the CorDis project, which examines, from an interactive discourse perspective, how the conflict in Iraq was discussed and reported in the Senate and Parliament and in various media outlets (Haarman and Lombardo 2009; Morley and Bailey 2009).

Example 4.6 CADS – dialogistic positioning

Duguid (2007), for example, examines the dialogistic positioning of Tony Blair and his two advisors in the Hutton inquiry. This was a British judicial inquiry chaired by Lord Hutton into the circumstances surrounding the death of Dr David Kelly, who had been named as the source of quotes

(*continued*)

by journalists saying that Tony Blair's Labour government had knowingly 'sexed up' a report into Iraq and weapons of mass destruction.

Duguid notes the frequency of the collective noun *people* in Tony Blair's evidence, which Fairclough (2000) has also noted surfacing as a keyword in his corpus of Blair speeches. In the corpus extracts below, the use of *people* serves to make the interactive, dialogistic nature of the discussions explicit, illustrating '… the continuous inter-textual concerns of the team, where a constant second-guessing goes on about how actions or texts will be perceived by those outside'.

> You should not have gone to war – *people* can have a disagreement about that … to, as it were, offer the name, but on the other hand, not to mislead *people* but *people* would say, 'when did you know?'
>
> (Duguid 2007: 91)

Studies in the CADS mould following the tradition of Fairclough's approach often apply Hallidayan grammar for linguistic analysis (see Section 3.2). See also Stubbs (1996) and Fairclough (2000) for CDA studies which utilise transitivity and nominalisations, respectively, in their analyses.

Corpus-informed critical discourse studies

Another perspective on corpus-based approaches to CDA, derived from the discourse-historical approach, is offered by the team of linguists working at Lancaster University on the project *Discourses of Refugees and Asylum Seekers in the UK Press 1996–2006*. Their research is based on the analysis of a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants and migrants (collectively referred to as RASIM).

A key difference between CADS and this discourse-historical inspired study is that in the RASIM project a wide spectrum of background information on the social, political, historical and cultural context of the corpus data was used both to formulate hypotheses on which to base research questions and also to inform interpretation of the corpus data. Key terms, e.g. 'refugee', were examined to see how they were conceptualised by 'official' sources such as dictionaries and organisations directly involved with these groups. Text-based analyses were also supported by official statistical information on the number of asylum applications.

Not only does this critical study pay attention to synchronic variation through an analysis of keywords, such as *refugee* and *asylum seeker*, occurring in both tabloids and broadsheets, but is also concerned with diachronic change, noting a causal link between events and corpus findings.

Concept 4.6 Causal link between events and corpus findings

> Clearly, major wars, natural disasters, and terrorist attacks resulted in an increased focus on RASIM [*refugees, asylum seekers, immigrants and migrants*]. However it is also interesting that two of the 'spikes' in our data co-occurred with political events in the United Kingdom: the Asylum Bill (March–April 2004) and the UK general elections (March–May 2005). During these periods, the construction of RASIM worked as part of (usually negative) media comment on government policies (what could be termed a 'political rivalry' discourse). At these times RASIM were thus functionalized as part of a struggle for political hegemony, being discursively constructed as a people who merely constitute the topic of political debate, somewhat dehumanised as an 'issue'.
>
> (Baker et al. 2008: 18)

The RASIM research has some affinity with the CADS approach with its focus on identification of keywords and collocation patterns, and their underlying semantic preference and discourse prosodies (Baker and McEnery 2005; Baker et al. 2008), but with less focus on SFL categories for linguistic analysis. It also differs from CADS in that these patterns were then mapped onto the discourse-historical CDA notions of *topos*, topic, and also metaphors commonly employed in racist discourse as a means of revealing elements of the underlying discourses relating to RASIM. For example, one of the common metaphors found to frame refugees was that of 'water', symbolising the loss of control over immigration, e.g. *immigrants are flooding the country*. However, Baker (2006) also advises supplementing quantitative studies with a more qualitative close reading of the text, so as to reveal other metaphors (e.g. refugees as invaders), which, *because* they are so infrequent, may be salient (see Charteris-Black 2004 for an exposition of his discourse model on critical metaphor analysis which considers the interdependency of semantics, pragmatics and also a cognitive dimension).

Corpus studies underpinned by the discourse-historical approach are few and far between, no doubt one reason being the intricate nature of the analyses drawing on a web of contextual strands at various stages of such a study as the one above.

Several issues regarding interpretation in corpus-based CDA studies have been aired in the literature. While corpus data may be used to examine different ideologies and sites of conflict and contestation, the interpretation of the corpus data may well be a site of contestation in itself. One way around this would be to involve the producers of the text in the interpretation, as Ward (2004) did by checking on the degree of inclusiveness of *we* in a corpus of union negotiations. Another way, though not always possible, would be to

allow 'multiple voices' in the interpretation: 'The authority, plausibility and reliability of the analysis can be further enhanced if the team members come from varied disciplinary backgrounds and bring diverse conceptual worlds and analytical tools to bear on the discourse' (Hardt-Mautner 1995: 4). This perspective is very much in line with the historical discourse approach to CDA, exemplified by the multifaceted approach of the RASIM project on the discursive constructions of refugees and asylum seekers.

Quote 4.6 A multidimensional CDA analysis

> Corpus analysis does not usually take into account the social, political, historical, and cultural context of the data. For this reason, a multidimensional CDA analysis that also goes beyond the 'linguistic' elements of the text is instrumental in allowing researchers to consider issues such as:
>
> - processes of text production and reception of the news data under analysis;
> - the social context of the news industry in the United Kingdom (e.g. the competitive news market);
> - (changing) political policy in the United Kingdom and elsewhere surrounding RASIM;
> - statistics regarding immigration and asylum applications;
> - social attitudes toward RASIM;
> - meta-data, e.g., reports or talk about the newspaper texts under examination;
> - macro-textual structures;
> - text-inherent structures (coherence and cohesion devices).
>
> (Baker et al. 2008: 33)

Aside from these difficulties regarding interpretation by the analyst, we also have to consider the effects of reception and how the text is likely to be interpreted by readers. In this respect, Stubbs (cited in Hunston 2001) notes that in educational discourse *falling standards* has become a fixed phrase, making it likely that people will come to regard standards as less high now than they previously were. Another example of a fixed phrase has been noted in their UK newspaper corpus by McEnery and Wodak, who found that *illegal asylum seekers* was a frozen collocate, but not the phrase *illegal refugees* (McEnery 2007). Would this mean that *asylum seekers* now has a different ideological meaning from *refugees* in the public's eyes? This question is not easy to answer and illustrates just how difficult it can be to establish ideological significance even on

the basis of seemingly more objective quantifiable corpus evidence. As Candlin (2007) points out, it would be necessary to have recourse to social psychology to ascertain the effects of reception and the impact of recurring patterns on individual readers.

Widdowson (2004: 89–111) has criticised CDA analysis for the fact that it 'collapses semantics and pragmatics: pragmatics is, in fact, reduced to semantics' (see Quote 1.8). Another key criticism is that researchers choose texts based on their own ideological stance: 'Texts are found to have a certain ideological meaning that is forced upon the reader …' (Blommaert 2005: 32). As Stubbs (2001b) and Mautner (2009a and b) have argued, though, recurrent patternings in large corpora may reveal certain ideological underpinnings, as demonstrated by O'Halloran and Coffin's research. This position is also supported by Kandil and Belcher (2011), whose research examined how the Israeli–Palestinian conflict was handled in three different news sources, Al-Jazeera, BBC and CNN. (It has to be borne in mind, though, that in some cases, saliency can take precedence over frequency in establishing ideological significance.) Extrapolating ideological meaning from corpus findings can be strengthened by involving producers of the text in the interpretation of data, involving researchers from different disciplinary backgrounds, instead of relying solely on the analyst's perception, and adopting a multidimensional analysis (see Quote 4.6).

4.4 Discourse analysis: spoken

Stenström's (1994) model of the exchange sequence of questions and responses, based on the Sinclair/Coulthard model, is a landmark in corpus analysis of spoken discourse. More recently, analysis of spoken discourse has largely concentrated on how various pragmatic devices and prosodic features are operationalised by participants in specific social situations.

4.4.1 Prosodic approach

Altenberg's (1998) pioneering work on phraseologies in the 500,000-word London-Lund Corpus (LLC) of spoken English of casual conversation has already been commented on in Section 3.1.2. Since then other studies making use of the London-Lund Corpus have adopted a more multilayered approach, paying attention not only to lexical, syntactic and discourse features, but also to prosodic elements (e.g. intonation choices, stress, pauses) of the discourse. Aijmer's (2002) research on the LLC is a prime example of this more finely grained prosodic approach, as is Wichmann's (2004) study of *please*-requests in the ICE-GB corpus (the British contribution to the International Corpus of English).

Example 4.7

In the LLC, out of 426 examples of *actually*, 217 of these were found in the first category in Table 4.2, 96 of which occurred in final position in the utterance.

Table 4.2 Aijmer's analysis of the discourse particle *actually*

	Initial (I)	**Mid (M)**	**Final (F)**
FACE (surreptitious)	32	89	96
FACE (non-surreptitious)	10	29	39
TELEPHONE	5	18	27
DISCUSSION	9	43	15
PUBLIC	1	8	3
PREPARED	–	3	–
Total	57	189	180

Aijmer follows Brazil's (1995) prosodic model of proclaiming and referring tones, according to which a tone unit consisting of fall-plus-rise is classified as a 'referring tone', as it contains information judged by the speaker to be already present in the context shared with the speaker and is therefore available to be referred to. With respect to *actually*, Aijmer notes that there seem to be several correlations between prosody and the discourse functions of *actually* in different positions in the utterance. For example, when *actually* is in final position the pattern fall-plus-rise tone has a social function indicating that the interlocutors share common ground. But where initial *actually* has a fall-plus-rise tone, this prosody is used for marking contrast.

(Aijmer 2002: 262)

Other researchers drawing on Brazil's prosodic model are Cheng and Warren who have prosodically transcribed a 2-million-word intercultural corpus of Hong Kong spoken English, the largest prosodically transcribed corpus currently in existence, marked up for prosodic features of prominence, tone, key and termination (Cheng et al. 2008b). Another unique feature of this corpus is that it is the first corpus-based study which maps speakers' discourse intonation choices onto word association patterns in terms of their collocational and colligational profiles. Cheng and Warren (2008) illustrate how prosodic choices can be exploited by speakers such that a politician, through use of a referring tone, might assert common ground where none exists.

4.4.2 Rhetorical approach

Various rhetorical features have been investigated in corpora of speech events in the academy, with the Michigan Corpus of Academic Spoken English

(MICASE) providing a wealth of information on key rhetorical devices deployed in spoken academic discourse.

It was noted in Section 4.3.3 that metadiscourse has been a major area of investigation in various written academic genres, with Hyland's work a pioneering endeavour in the field. Metadiscourse has also been extensively studied by various researchers using MICASE data on academic lectures, most notably by Mauranen (2001, 2003a, 2004b). Whereas Hyland views metadiscourse as essentially interpersonal, Mauranen, on the other hand, accords it a mainly textual role, but also acknowledges its interpersonal dimension.

Concept 4.7 Mauranen on metadiscourse in spoken academic genres

> As metadiscourse organizes ongoing discourse (*so let me just elaborate a little bit and then we* …), it is obviously fundamentally textual. The textual process of organising discourse as it unfolds also imposes the speaker's order on the discourse situation, and in this sense can be seen as acting out power relations. This in turn relates to the interpersonal function, as is clearly seen in the evaluative modification of metadiscourse and the concomitant labeling of speech acts (*that's a very interesting remark/good observation/a really important point*), which reproduces relations of power … and … socialize students into the discourse community.
>
> (Mauranen 2003a: 21–2)

Other research on the MICASE data reveals how extensively metadiscourse occurs with various types of hedging devices, another linguistic phenomenon researched in depth by Hyland (1998) (see Concept 4.3).

Quote 4.7 Hyland's definition of hedges

> Hedges are devices which withhold complete commitment to a proposition, allowing information to be presented as an opinion rather than fact (Hyland 1998). They imply that a claim is based upon plausible reasoning rather than certain knowledge and so both indicate the degree of confidence it might be wise to attribute to a claim while allowing writers to open a discursive space for readers to dispute interpretations.
>
> (Hyland 2009: 75)

Mauranen (2004b), in fact, distinguishes between two kinds of hedges, epistemic and strategic, with Mauranen's definition of epistemic hedges as 'indicating conceptual openness' echoing Hyland's opening of a discursive space.

Example 4.8 Mauranen on strategic and epistemic hedges in MICASE

Expressions such as *a little bit* or *just* incorporated into reflexive discourse are mainly used for softening and mitigation, i.e. used strategically:

> but again *I just want to try to get across* the general issue that we're interested.
>
> (Lindemann and Mauranen 2001: 5)

Phrases such as *kind of*, *sort of* and *or something* have more epistemic uses, indicating conceptual openness, affecting the propositional content of a statement:

> But now with this factor, the m- ma- the scope is maximized, at a *somewhat*, lower temperature.
>
> (Mauranen 2004b: 174)

Quite often, both types of hedges combine together in metadiscourse:

> *Let me just … rephrase this … let's just write it up here or something* so you can, take (note).
>
> (Lindemann and Mauranen 2001: 5)

Specifically, Mauranen (2004b) notes that these two different types of hedges, strategic and epistemic, tend to be genre-specific, with the more dialogic genres (e.g. tutorials) found to have more strategic hedges and the monologic genre of lectures more epistemic hedges. Moreover, certain hedges, e.g. *sorta*/*kinda*, were found to be commoner in the social sciences and the humanities lectures, and not necessarily equated with female speech as usually claimed (Poos and Simpson 2002). *Sort of*, *kind of* and *you know*, have also been investigated in the BASE (British Academic Spoken English) lecture corpus, the British counterpart to MICASE, by Lin (2010). She highlights the multifunctionality of what she terms 'pragmatic force modifiers', noting that they 'guide students to intended inferences, present a cautious attitude towards expressing opinions and create an ambience of friendliness and casualness simultaneously' (p. 1180).

To sum up, explorations of MICASE and BASE mark a significant advance in the field of corpus-based enquiries of academic spoken discourse. Several of the mitigating expressions such as *just* and *sort of* had already been researched in the London-Lund Corpus in the 1980s and early 1990s, and were analysed in respect of their occurrences in monologue and dialogue. However, the detailed mark-up of

speech categories for the MICASE data including participants' age, gender, status, etc. make for much more finely tuned analyses sensitive to the sociocultural context of the data, as is also the case with the studies of CANCODE (see Section 3.4) and those on the Hong Kong Corpus of Spoken English discussed earlier.

The above exposition of corpus-based studies of various spoken and written genres from a discourse-analytic perspective has revealed just how closely language choice is affiliated with the epistemologies of different academic disciplines and the role that disciplinary practices play in shaping a genre. Corpus-based analyses have also revealed the interactive nature of discourse and uncovered the complex interplay of propositional and interpersonal elements in meaning-making, the true intentions and purposes of which can only sometimes be fully and faithfully interpreted by bringing a more ethnographic perspective to bear on the analysis.

4.5 Discourse analysis: multimodal

Another way in which the field of corpus linguistics is branching out is to take a more multimodal rather than a purely monomodal approach to corpus analysis. Such corpora fall into roughly four types. Some purely audio-visual corpora, such as the Scottish Corpus of Texts and Speech (SCOTS), can be searched according to marked-up sociolinguistic variables such as age, sex, etc. (see Sections 5.4.3 and 7.4.5). Other types of multimodal corpora, on the other hand, are now being compiled with a view to establishing interrelationships between the linguistic and non-linguistic aspects of discourse, thereby taking up Gee's (2001) notion of Discourses, involving not only linguistic practices but other semiotic elements (see Quote 4.3). In the following subsections, three main undertakings from different theoretical-analytic approaches, which all seek to analyse the verbal and visual at a discourse-based level, are discussed.

4.5.1 SFL approach

Pioneers in this field working in the Hallidayan tradition (Baldry and Thibault 2001, 2006) take a systemic-functional orientation to multimodal corpus linguistics to determine how different semiotic resources (language, gaze, gesture, etc.) interact to create meaning.

Concept 4.8 Multimodal corpus linguistics

Baldry and Thibault (2001: 87) argue that the pre-eminent role of language in meaning-making is unwarranted and that other semiotic resources, such as gesture, gaze, colour, movement and voice quality should also be taken into account: 'such an approach will help us to understand the textual and contextual principles governing the co-deployment of meaning-making resources'.

In keeping with this meaning-oriented approach, Baldry and Thibault have devised a functionally oriented concordancer and also a tagging system for analysing both linguistic and visual modalities, such as transitivity frames, in the system networks. This is a major departure from previous work in tagging and searching a corpus, which relies on purely formal criteria, e.g. the linguistic environment in which a particular lemma occurs, for exploring a corpus.

Example 4.9 Visual semiotic: gaze transitivity frame

In a multimodal corpus of car advertisements, the car is often found as the Phenomenon in a gaze transitivity frame, as illustrated below. Baldry and Thibault (2006) have proposed the notion of 'visual collocation' to refer to the probability of co-occurrence of constellations of visual items in a particular setting. For example, in car advertisements featuring test drivers, the cars in question are found to collocate with difficult testing terrains, such as deserts (Table 4.3).

Table 4.3 Visual collocation

	Participant	**Process**	**Circumstances**
Text 1	Gazer: Young adult male: driver; Phenomenon: row of Honda cars	Right–left gaze vector extending from eyes of Gazer to implied Phenomenon; Right–left head turn establishing direction of gaze vector; Inter-shot cut extends vector from Gazer to Phenomenon	(i) Location: driver beside Honda Solar car; (ii) Manner: smiling (Gazer); (iii) Location: line of cars in desert scene

(Baldry and Thibault 2006: 173).

Baldry and Thibault's ultimate objective is to develop a theoretical model which can account for non-linguistic semiotic resources as meaning-making systems in combination with language resources, which can be extended to a wide range of multimodal corpora.

4.5.2 Functional approach

A different kind of multimodal analysis is that carried out from a functional perspective in the *HeadTalk* project at Nottingham University. One of the initial purposes of this project, making use of videotaped MA and PhD supervisions, is to explore the alignment of non-verbal backchannelling head nods with co-occurring verbal backchannels (e.g. *yeah, right, mm*) in order to uncover new

understandings of textuality (Adolphs and Carter 2007; Carter and Adolphs 2008). A sophisticated coding scheme has been devised for defining and classifying both the verbal and non-verbal backchannels, as outlined below.

Concept 4.9 Coding scheme for verbal and nonverbal backchannels

Framework for coding key functions

- continuers: maintaining the flow of discourse
- convergence tokens: marking agreement and disagreement
- engaged response tokens: high level of engagement, with the participant responding on an affective level to the interlocutor
- information receipt tokens: marking points of the conversation where adequate information has been received

Framework for coding head nods

Five broad types of head nods:

Type A: small (low amplitude) nods with short duration
Type B: small (low amplitude) multiple nods with a longer duration than type 1
Type C: intense (high amplitude) nods with short duration
Type D: intense and multiple nods with a longer duration than type 3
Type E: multiple nods, comprising a combination of types 1 and 3, with a longer duration than types 1 and 3.

(Adapted from Carter and Adolphs 2008: 279–81)

Carter and Adolphs (2008) emphasise the discourse-level perspective of their analysis, pointing out that it is 'critically important […] that corpus-based approaches to text engage with the level of discourse analysis and discourse-level meaning relations on various scalar levels of textual organisation'. The project is still ongoing and Carter and Adolphs suggest several possible avenues for future exploration, e.g. Do gestures have a syntax, that is, are they syntagmatically and/or paradigmatically organized? (p. 287).

4.5.3 Situated discourse approach

Another pioneering endeavour in the field, and perhaps the most ambitious, is that being undertaken by Yueguo Gu under the auspices of the Chinese Academy of Social Sciences of the compilation of a Spoken Chinese Corpus of Situated Discourse (SCCSD (Gu 2002, 2006). In this situated discourse

approach the orthographic transcription takes a supplementary role in the data description and analysis, with the primary role being played by video streams and synchronised sounds. Gu's work has some affinity with Halliday's systemic-functional grammar, a system in which language is viewed as a social phenomenon and used for meaning-making. The model also draws on Kress's (2001) work on multimodality in its study of social action over space and time. Gu (2006: 138) regards this model as an interface bridging society and the individual, 'where the macro, a social system, and the micro, the individuals within the system, interact'.

The model aims to encapsulate the 'real world' nature of situated discourse, including its 'situatedness' to an actual situation, to actual users, to actual goals, to spatial and temporal setting, and to the cognitive capacity of actual users. Likewise, the talking and doing level of the model are interwoven in 'real world' situations.

Concept 4.10 Real world situated talking and doing

1. Talking is the task, e.g. meeting, seminar (it is task-oriented, task-goal-directed);
2. Talking is the main constitutive part of the task, some classroom discourse, doctor–patient discourse (it is task-oriented, task-goal-directed);
3. Talking is a constitutive part of the task, e.g. giving instructions from time to time (task performance is dominant, talking tends to be fragmented);
4. Talking and doing run in conflicting parallel, the achievement of the latter serves as a means to the goal of the former, e.g. business dinner (business table talk);
5. Talking is an embedded social part of the task, e.g. talking over the meal (talking has no specific goal to reach);
6. Talking is a decorative part of the task, e.g. talking accompanying tea-making;
7. Talking is a hindrance to the task, e.g. talking over a written exam;
8. Talking and task are independent of each other.

(Adapted from Gu http://ling.cass.cn/dangdai/gu_papers/)

The groundbreaking research studies outlined above demand, by necessity, new software and tools for examining discourse from this multifaceted perspective. This research also, in turn, fosters new cross-disciplinary alliances. What is of interest to note, though, regarding these three major initiatives in analysing multimodal corpora is that they are all confined to individual institutions. It is therefore expected that when these projects have matured, the tools and

software will 'migrate' to other corners of the globe to further advance the relatively new but rapidly developing field of multimodal corpus linguistics.

4.6 Discourse analysis: hybridisation of modes

There is now a 'new modal order' emerging in this era of digital literacies, specifically computer-mediated communication (CMC) involving e-mail, discussion groups, Internet relay chats (IRCs) and weblogs, etc., entailing a complex hybridisation of spoken and written modes. Although some analyses have been carried out into these 'new technologies' corpora, mainly of weblogs (Ooi et al. 2007, Ooi 2009), this is still a fledgling area as far as corpus-based discourse analysis is concerned and one which poses enormous challenges for existing software.

Example 4.10 Challenges in using part-of-speech taggers to handle CMC features

Ooi (2009) reports on a study using the CLAWS tagger (Garside et al. 1997) to analyse a short conversational exchange in IRC:

> ** Now talking in #cybercafe**
> **archangel has joined #cybercafe**
> <archangel> *wat* are u guys *tokking* about?
> <Cyclops> Iris is talking about her experiences at her mentoring sessions …
> <archangel> *oic*, pls continue …
> <Cutie Pie> as I was saying, my school does mentoring in groups … *b4* the first session, we split ourselves into groups of guys and gals, instead of the *teech*, doing *4* us …

Ooi found the following features posed a challenge to CLAWS: *wat* = 'what'; *tokking* = 'talking'; *oic* = 'Oh, I see'; *b4* = 'before'; *teech* = 'teacher'.

(Adapted from Ooi 2009: 112)

Not only will new software be required to analyse shortened forms and new spellings, as in the above example, but also other discourse features of Internet communication such as emoticons, which can have evaluative, expressive or regulative functions, and other conventions, such as upper case, for simulating prosodic features. Moreover, King (2009) notes the challenge for corpus linguists in analysing turn-taking in chat rooms in which one turn can often be split

into many in order to keep up with the real-time unfolding of conversation. And how can all the semiotic elements in a corpus of weblogs with its multifold modalities (text, video, pictures, audio files, hypertextual links to other blogs) be accommodated within a discourse-analytic framework?

4.7 Conclusion

This chapter on corpus-based discourse analysis has exemplified how the field has moved from single-pronged to more multi-pronged approaches, from a language-in-use to a more complex language-in-action perspective, and from monomodal to multimodal analyses (see Quote 4.3). This complex synergy of methods, approaches and tools has enabled a rapprochement of the two fields, with corpus linguistics no longer hovering on the periphery of discourse analysis but now assuming a central role. However, new forms of discourse are evolving which have thrown up new challenges for both software developers and corpus analysts.

Further reading

Adolphs, S. (2008) *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins. This monograph examines various speech acts in the CANCODE corpus.

Aijmer, K. and Stenström, A.-B. (eds) (2004) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins. This edited collection contains a range of corpus-based studies on different aspects of discourse: cohesion and coherence, meta-discourse and discourse markers, and text and information structure.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum. This book provides a very reader-friendly introduction on how keywords, frequency and dispersion and collocational networks can inform discourse analysis, especially those studies of a CDA nature.

Grundmann, R. and Khrishnamurthy, R. (2010) The discourse of climate change: a corpus-based approach. *Critical Approaches to Discourse Analysis across Disciplines*, 4 (2): 125–46. This article analyses data on climate change, drawn from US, UK, French and German sources in Nexis, a database of global news sources.

Paltridge, B. (2006) *Discourse Analysis*. London: Continuum. Chapter 7 in this volume has a very useful overview of corpus approaches to discourse analysis.

5
How is Corpus Linguistics Related to Sociolinguistics?

This chapter will:

- Discuss the two main approaches from sociolinguistics for analysing corpus data
- Show how sociolinguistic analyses intersect with discourse analysis, especially in the area of pragmatic markers
- Outline the limitations of sociolinguistically motivated corpus analyses

Section 3.4 has presented the sociolinguistic approach of the Nottingham school to analysis of the CANCODE corpus. This chapter seeks to give a broader perspective on the interrelationship between corpus linguistics and sociolinguistics with reference to key studies, the majority of which focus on spoken language. This chapter also takes up the key theme of Part II 'The Nexus of Corpus Linguistics, Textlinguistics and Sociolinguistics' as the ensuing subsections reveal how many of these analyses are also discourse-based, especially with their focus on pragmatic markers of various kinds.

5.1 Definition of sociolinguistics

Quote 5.1 Definition of sociolinguistics

> Sociolinguistics is the branch of linguistics that is concerned with linguistic variation and its social significance. Sociolinguists study differences between individual users of language and language varieties. The discipline encompasses many subtopics which focus on the effect which various social characteristics have on the language of individuals or speaker groups. Sociolinguistic factors include the major demographic categories of age, gender, social class and ethnicity, as well as situational categories like the degree of formality of the speech situation, the social networks of the speaker and so on.
>
> (*continued*)

> Sociolinguistics also focuses on the overall characteristics of language varieties: that is regional dialects, standard/non-standard varieties of a language, multilingualism, language policy, standardization etc.
>
> (Andersen 2010: 547)

There exist two main approaches to sociolinguistic research, 'interactional sociolinguistics' and 'variational sociolinguistics', although, in effect, much corpus-based research draws on both approaches.

Concept 5.1 Research approaches to sociolinguistics

In the variationist approach, the language use of a particular group is analysed quantitatively according to particular linguistic variables, which can be phonetic, prosodic, lexical, grammatical or pragmatic. These variables are mapped against demographic data such as age, gender, social class, ethnicity, geographic location, etc. Research in the variationist paradigm is often diachronic in nature, charting language change.

In the interactional approach more qualitative and ethnographic methods are employed to study communicative interaction in various speech situations. This approach views language as an unfolding dynamic entity, co-constructed in the process of interaction, and draws on conversational analysis and pragmatics for the situational analysis.

5.2 Corpus studies in the interactional paradigm

It is clear that analysis of the CANCODE data exemplified and discussed in Section 3.4 is mainly situated within the interactional paradigm. Other corpus studies following the finely grained, qualitative sociolinguistic approach of the Nottingham school are those by O'Keeffe (2006) using a corpus of media discourse, and Koester (2006, 2010) on workplace discourse, in part based on the 1-million-word Cambridge and Nottingham Business English Corpus (CANBEC), a specialised corpus closely modelled on CANCODE.

5.2.1 Conversation analysis perspective

Koester notes the similarity of interactional sociolinguistics to conversation analysis in which talk is seen as creating its own context, where context is defined as 'dynamically created and expressed in and through interaction', i.e. context is viewed as 'talk-intrinsic' (Koester ibid.: 12). 'Contextualisation clues' are used by speakers and listeners, collaboratively, to signal and make inferences about communication goals (Gumperz 1992). Such contextualisation clues can operate at the level of individual speech acts or at a more global level where the notion of 'frames' (Goffman 1974) indicates generic activity, and the context is viewed as 'talk-extrinsic'.

Example 5.1 Decision-making frame in workplace communication

Koester (2006: 36) notes that in her corpus of workplace communication decision-making frames often begin with the identification of some kind of problem. In the data from the back office of an American food cooperative, the bookkeeper begins a conversation with her co-workers as follows:

Ann: Anyone wanna decipher some handwriting?

Koester remarks that the word *decipher* signals a decision-making frame, as it indicates the existence of a problem. She also remarks on the interactive nature of this genre, signalled by the formulation of the initial turn as a question (a request or offer), thus underscoring Gumperz's concept of speech as a collaborative, coordinated activity.

O'Keeffe (ibid.) also establishes a link between her data on routinised opening formats of radio phone-ins and canonical openings in conversation analysis, similar in concept to Goffman's 'frames'.

Example 5.2 Opening sequences of radio phone-ins

O'Keeffe (2006) remarks on the similarity between canonical opening sequences associated with conversation analysis and the opening formats of radio phone-ins. However, her corpus data show openings of radio phone-ins to be more elaborated, involving a collaborative setting up of the reason for calling. She also notes the use of display questions in such sequences, where the presenter poses a question to which the answer is known (e.g. *You are attending boarding school at the moment?*), as in the example below.

Extract 3.1.2

1. Presenter: Welcome back to the program Aidan good afternoon to you. Hello there.
2. Caller: Hello.
3. Presenter: Hi how are you?
4. Caller: Fine thanks.
5. Presenter: You are attending boarding school at the moment?
6. Caller: Yeah I'm in my first year at the moment I'm just on a break as I've just gotten I'm feeling I've got a chest infection so I'm at home at the moment.

(Adapted from O'Keeffe 2006: 56–7)

5.2.2 Ethnographic perspective

Although identification of routinised ways of saying things may, in turn, help in the identification of speakers' goals and intentions, this may not always be the case. In some instances a more ethnographic approach may be in order, in which aspects such as institutional practices and the broader and local social organisational conditions play a central role in analysis of the data. Conversation analysts do not take these contextual features into account for analysis of the dialogic data. Koester (2006) shows just how crucial such considerations can be for interpretation of the data, especially in the case of speech acts, the interpretation of which very often relies on knowledge of social contexts and social relationships (see Section 4.3.1 on genre approaches to discourse analysis). Without access to such contextual information the extract cited below would make little sense.

Concept 5.2 Ethnographic analysis of workplace conversation

Koester (2006) notes that for interpretation of a corpus of workplace discourse it is important to know both the institutional and immediate context and also the physical context. In the extract below we need to know that the conversation takes place in a printing firm and that Dave is the firm's artist and Val is responsible for putting together quotations for customers. The nature of their relationship is also of importance for interpreting the discourse. The fact that Val and Dave are on the same level within the company and have a friendly relationship means that turn 9 (*'cause you haven't got a clue what you're doing*) has a bantering rather than a threatening tone. The physical documents that Dave is showing Val are also to be considered as an integral part of the interaction.

Example 2.1

1. Dave: [And she said something about Christmas brochures, this was on Friday, I can't remember what it was, ↑ Is she waiting for a quote on *that* still?
2. Val: Christmas brochures … ↑ Uhm…
3. Dave: Ah! … Oh I see.
4. Val: Is it – ↑ Are these the ones that we're doing. That they want −
5. Dave: to go in, ↑ Oh did they want a separate quote on that, I suppose, =
6. Val: = Which. Which ones. Which *ones*.
7. Dave: You know the … four page things? Have you not seen it. It's the same style but from Malcolm Hennessy. No it's not Malcolm Hennessy, it's Delaney.
8. Val: [Delaney

9. Dave: Christmas … So it would be … I have to refer to her again, ↑ 'Cause *they* wanna come through *you* really, 'cause *you* haven't got a clue what you're *do*ing.
10. Val: No.

(Adapted from Koester 2006: 13)

Cutting (2008: 93–100), meanwhile, discusses corpus analysis from the perspective of communities of practice, specifically the phraseology of the discourse marker 'right' in a corpus of casual conversations from a student community. She exemplifies how knowledge of the corpus as a whole and the importance of corpus mark-up for interlocutors can shed light on the analysis. For instance, one interlocutor was found to use 'that's right' and 'right' far more frequently than the others; Cutting suggests that this may be because this student listens more than he talks, he is more of a solidarity-giver, or it may just be part of his idiolect. Nevertheless, these multiple interpretations are a reminder of how just how difficult it is to ascribe reasons for certain language use. By way of example, Rayson et al. (1997) found from their analysis of the BNC data that people from socially disadvantaged groups used more taboo terms (e.g. *bloody*) than people from advantaged groups, but the reason for this (e.g. upbringing, use of terms to show group solidarity) is ultimately based on the analyst's own interpretation, possibly influenced by their own biases and identities.

In some ways, dividing corpus research into interactional and variationist approaches is somewhat of a false dichotomy as, in reality, corpus studies draw on aspects of both. One such study is that by He and Kennedy (1999), who looked at successful turn bidding, i.e. interruptions, across three speech domains in the LLC (casual conversation including private discussions and interviews, public discussions including radio discussions and meetings, and telephone conversations including private and business conversations). These successful turn biddings were analysed according to sociolinguistic variables such as the level of familiarity between interlocutors, their relative status, and gender, which were found to be significant factors in the hearer's choice of a boundary (i.e. a prosodic tone unit ending with a silent pause, or a lexical boundary in the form of an address tag or emphatic marker) for initiating a successful turn. Corpus studies of a mainly variationist nature are discussed below.

5.3 Corpus studies in the variationist paradigm

Those corpus studies taking a purely variationist approach mainly centre on dialectology and varieties of English, investigations which, by their very nature, analyse language primarily according to geographic location and ethnicity, but sometimes also with reference to other sociolinguistic variables such as age,

gender and social class. Other corpus studies usually considered as variationist include those of a contrastive diachronic nature, mapping language change over different time periods.

5.3.1 Dialect corpora

Concept 5.3 Definition of 'dialect'

> A dialect is not a distinct language as such, but a variety of a language spoken in a particular geographical area of a country, for example 'Geordie' English (from Newcastle upon Tyne in the UK) and 'Estuary' English widely spoken in south-east England, especially along the River Thames and its estuary. Dialects can also be defined according to social class, with Estuary English found to be increasingly used by speakers of various social classes.

Corpus work on dialectology has been somewhat sparse with regard to recent corpora compared with the extensive research on regional dialects in Old and Middle English (850–1710) in the 1.5 million-word Helsinki Corpus (Rissanen 2009). One reason could be that it is perceived as less meaningful than corpus work carried out on different varieties of a language or different languages (McEnery et al. 2006). Nevertheless, several noteworthy studies exploring dialect in more modern-day speech corpora do exist.

One corpus, the Survey of English Dialects (SED), was started in 1948 by Harold Orton at the University of Leeds. Recordings from 318 locations all over rural England were made, supplemented with extensive interviews. The recordings, which were made during 1948 to 1973, consist of about 60 hours of dialogue of elderly people talking about life, work and recreation (Orton 1962). A more recent dialect corpus is the IViE (Intonational Variation in English) corpus which contains recordings (approximately 40 hours) of 9 urban dialects of English spoken in the British Isles. In one of the associated projects a computational model of intonation was constructed which took account of variation due to dialect, speaking style, gender and individual speaker habits (http://www.phon.ox.ac.uk/IViE/). Smaller-scale, more regionally focused dialect corpora have also been compiled. One British English dialect corpus is the 1.5 million-word corpus of York English, analysed for social and regional dialect patterns (Tagliamonte and Lawrence 2000).

5.3.2 Varieties of English corpora

In looking at how corpus linguistics has been applied to different varieties of language, it is useful to discuss this issue within the framework of Kachru's (1986) 'World Englishes' tradition, as major corpus initiatives adopt this macroscopic framework.

Concept 5.4 Language variety and 'World Englishes'

Language variety is a more general term than 'dialect', which is specifically designated for regional varieties. Language variety refers to a 'variant of a language that differs from another variant of the same language systematically and coherently' (McEnery et al. 2006: 90). Varieties of English are often referred to as 'World Englishes' (cf. Bolton and Kachru 2006), and considered as belonging to either 'inner' or 'outer' circle varieties:

1. *The inner circle*, which contains the native English-speaking countries (UK, USA, Australia, etc.).
2. *The outer circle*, which contains former colonies of the UK and USA (India, Kenya, Nigeria, Hong Kong, etc.). These countries have developed nativised varieties of English which have achieved the status of official language and/or language of education, administration, etc.

Corpus studies of 'inner' circle of varieties of English

Studies of 'inner' circle varieties of English, specifically British and American English, using the 1-million Lancaster-Oslo-Bergen (LOB) Corpus and its American counterpart, Brown, both dating from the 1960s, are already well documented. There also exist numerous reports on the updated versions of these two corpora representing English as used in 1991, the Freiburg-LOB Corpus of British English (FLOB) and the Freiburg-Brown Corpus of American English (FROWN), on which diachronic and synchronic studies have been carried out (see McEnery et al. 2006 for summaries of these).

For example, in one diachronic contrastive study Holmes and Sigley (2002) used Brown/LOB, FROWN/FLOB and the Wellington Corpus of New Zealand English (http://vuw.ac.nz/), combining the same basic categories as the aforementioned corpora, to track social change in patterns of gender-marking between 1961 and 1991. Interestingly, they attribute the double increase in the references to women and the trend in the use of forms suffixed or pre-modified by *woman* to a greater recognition of the existence of, for example, 'women judges', concluding that 'the marking increase thus reflects both increased real-world participation, and continued attention to equal opportunity issues' (p. 261). See Section 8.4 for a study looking at gender-marking from a social psychological perspective. Another corpus modelled on Brown and LOB, to enable direct interdialectal comparisons, is the Australian Corpus of English, ACE (see Green and Peters 1991).

One large-scale initiative is the International Corpus of English (ICE) project (see Greenbaum 1996 and http://www.ucl.ac.uk/english-usage/ice/), comprising inner circle varieties of English corpora, e.g. ICE-GB (Great Britain)

and ICE-NZ (New Zealand), in addition to outer circle varieties (see following section), comprising 12 different registers. Most importantly, all the 1-million-word ICE subcorpora have been collected in such a way so as to allow comparison of different varieties of world Englishes. Some such as ICE-GB are marked up with sociolinguistic information on speakers and writers, although this subcorpus has been criticised for the fact that it represents only educated London English. Although most studies using the ICE subcorpora focus on comparing an inner circle variety of English with an outer one, there are a few which focus on cross-comparisons of inner circle varieties. Peters' (1996) study uses data from the British and Australian components of the ICE corpus to examine aspects of the comparative clauses conjoined with correlative *than* and *as* and to compare their use in Britain and Australia. Another study comparing different varieties of inner circle English is that by Wong and Peters (2007), who carried out a two-way and three-way study of backchannels (feedback given while someone else is talking, e.g. *uh-huh*, to show interest, attention and/or a willingness to keep listening). See Section 4.5.2 for a study by Carter and Adolphs (2008) on backchannels using a multimodal corpus.

Example 5.3 Backchannels in New Zealand, Australian and US English

Two-way comparison
Comparing the use of backchannels in transcriptions of telephone conversations drawn from the Australian and New Zealand subcorpora of ICE, Wong and Peters (2007: 506) conclude:

> The large proportion of single backchannels in the Australian data suggest that Australian listeners are more likely to direct their support towards both conversational management and supporting the speaker's right to continue in the turn. The New Zealand data are less clearly divided along these lines. However, their inclination to produce large quantities of backchannel clusters suggests that New Zealand listeners are as likely to direct their support towards the content of the message as to the speaker. Overall, this suggests that Australian listeners are more oriented towards maintaining conversational continuity by waiting for their turn at the talk. The New Zealand data suggests a tension between the need to support other conversationalists and the need to assert the right to speak when desired.

Three-way comparison
Findings from the two-way analysis were then compared with backchannel usage by US English listeners drawn from White's (1989) data. This

comparison showed that while English listeners worldwide draw on a common repertoire of backchannel forms, comprising variants of *mm, yeah oh, mhm* and *uhuh*, they differed in the complexity of the structures they used. For example, it was found that whereas US listeners used this set mainly as single backchannel expressions, Australian and New Zealand listeners tended to use them as anchors for more complex backchannel clusters.

Corpus studies of 'outer' circle of varieties of English

ICE also comprises corpora reflecting the 'outer' circle of English, such as ICE-Jamaica (Mair 2009), ICE-Singapore (Ooi 1997) and ICE-Hong Kong (Bolt and Bolton 1996). The *ICAME Journal* (vol. 34, April 2010, 'ICE Age 2: ICE Corpora of New Englishes in the making') is devoted to articles reporting on the latest endeavours in compiling other ICE varieties, such as ICE-Malta and ICE-Trinidad and Tobago. It is now widely accepted that Singaporean English and Indian English exist as distinct varieties, evidence for which is supplied by corpus data. Ooi et al.'s (2007) research analyses the English of personal weblogs from Singaporean teens (see Section 4.6 on 'new technologies' corpora), finding differentiation in the linguistic styles of males and females, and items characteristic of online Singaporean English, e.g. the discourse particles *de, lah, sia*.

However, the picture is not so clear for Hong Kong English. Bolton (2003: 205) refers to two stages in Asian Englishes, recognition and legitimation, stating that: 'In Hong Kong, the first stage of recognition of Hong Kong English is still wanting among some academics, although survey results indicate that the general Hong Kong public are aware of the existence of a distinct local variety.' Steps towards establishing 'legitimation' of Hong Kong English through corpus linguistics evidence are outlined below.

Concept 5.5 'Legitimation' of Hong Kong English through corpus data

Legitimation concerns whether particular words which spring up usually to express key features of the physical and social environment and which are regarded as peculiar to the variety, have become so regularised that they are deemed acceptable by the community at large, rather than by some outside gatekeeping authority (Butler 1997).

Corpus data have established the existence of Hong Kong English for lexical items. For example, Bolton (2003) has drawn up a glossary of 318 words of Hong Kong English vocabulary (e.g. *sampan, tea money*) based on

(*continued*)

Asiacorp, a 10-million-word database made up of fiction, non-fiction and print media English language texts used as a resource in the Macquarie dictionary project.

However, even to this day, except for a few studies on modified grammatical forms (cf. Gisborne (2000) on the patterning of relative clauses), Hong Kong grammatical varieties remain under-researched, most probably for the reason that lexical items are more likely to be a marker for a regional variety of English. Localised corpora such as ICE-HK would therefore provide a good base for investigating recurring grammatical patternings, and, based on the text types in which they occur, determining whether these should be classified as interlanguage errors or a variety of English (Li 2000) for standardising and legitimating the recognition of Hong Kong syntactic features.

A key study comparing five ICE subcorpora, namely Great Britain (ICE-GB), Hong Kong (ICE-HK), India (ICE-IN), the Philippines (ICE-PH) and Singapore (ICE-SG), is that by Xiao (2009). In this rigorous study Xiao applied Biber's multidimensional analysis, which is mainly confined to grammatical categories (see Section 3.3), and enhanced this with semantic features tagged by the WMatrix software (see Rayson 2008). Statistical analyses showed there to be key differences among the five ICE components across different registers, which Xiao accounts for through colonial history, influence of the native language and language contact.

Example 5.4 Variation across five varieties of English

When plotted against Biber's dimension of 'interactive casual discourse vs. informative elaborate discourse' (see Section 3.3.1), Indian English was found to be less interactive but more elaborate in nearly all registers when compared with the other four varieties of English. Xiao offers the following explanation for this finding: 'This is partly a legacy of the Raj and the East India Company, and partly a result of influences of native Indian languages, which give primacy to verbs rather than nouns' (p. 443). In contrast to Indian English, British English was the most interactive and least elaborate in registers such as private and public conversations and instructional writing.

The three varieties of English as used in South East Asia (Hong Kong, Singapore and the Philippines) were found to be very similar along this dimension, lying between British English and Indian English. Xiao accounts for the similarity by the fact that Singapore and Hong Kong share a common background language of Chinese. Another explanation given is that these

> Asian varieties of English influence each other through language contact; for example, Philippine English has some influence on Hong Kong English because of the large number of Filipino domestic workers in Hong Kong.
>
> (Adapted from Xiao 2009: 442–3)

Although corpus linguistic techniques have been extremely valuable for identifying the inner and outer circle regional varieties of English, these two should not necessarily be regarded as static and dichotomous. Hundt and Biewer (2007: 250) present data on variation between the present perfect and past tense in a South Pacific and East Asian corpus (SPEAC) of newspapers collected from the Web (see Section 1.1.2). Their preliminary results suggest that with increased contact between East Asian countries and Australia and New Zealand, the dynamics of inner and outer circle varieties are changing, and 'inner circles might have become or be in the process of becoming new epicenters in the South Pacific and parts of East Asia; they function as a model for outer circle varieties like Philippine and Fiji or Singaporean English'. From these two studies, language contact, for whatever reason, seems to be having considerable influence on both inner and outer varieties of English of the Pacific Rim countries.

Corpus linguistics is thus coming to play an increasingly important role in the identification and codification of various features of dialects and inner and outer circle varieties of English (the 'expanding circle' of World Englishes is discussed in Section 6.1; see Concept 6.1). It is expected that the ICE project, an enormous undertaking still in progress, as the American and Fijian subcorpora are still to be completed, will continue to provide a wealth of material for future intra- and inter-varietal studies.

5.3.3 Variationist 'other languages' corpora

In the previous section corpus studies of English in the variationist paradigm were discussed from the perspective of 'World Englishes'. Studies of 'other languages' corpora can be considered variationist, either because they examine a language from different time periods, i.e. from a diachronic perspective, or because they examine different varieties of a language from a synchronic perspective. English, or the varieties hereof, have dominated the scene as regards the studies of diachronic corpora, which have usually been of an historical nature (see Rissanen 2009 for a review). The type of study carried out by Beeching (2006) on two corpora of French from different time periods, the (1966–70) Orléans Corpus and the (1980–90) Bristol Corpus, are few and far between.

Example 5.5 Diachronic variation in corpora of French

Beeching (2006: 56–7) notes that *hein* and *quoi* are mildly stigmatised as they do not occur in formal written French and would also be highly unlikely to occur in formal spoken French. Both can be considered as hedging devices serving social/interactional purposes, with *hein* generally translated into English by a tag question or by *you know*?, and *quoi* translated as *as it were*, *know what I mean*?:

Oui, peut-être mais ça depend aussi, **hein**?
Yes, perhaps, but it depends, too, doesn't it?

ne pas avoir que des contraintes dans dans sa vie, **quoi**, hein?
not just to have obligations in in one's life, as it were, you know?

Beeching's corpus data show that while rates of *hein* usage were similar overall in both corpora, the class distribution of usage differed dramatically with middle-class speakers adopting *hein*. Rates of *quoi* usage doubled in the later corpus, with much higher rates found among middle-class speakers. Beeching tentatively concludes that *hein* and *quoi* are becoming less stigmatised and that 'it seems possible that, if middle class speakers are beginning to adopt stigmatized "working class" speech forms, there has been a democratization, a shift in the hierarchical nature of French society' (p. 58).

The above example from Beeching (2006) examines hedging devices from a diachronic perspective (see Section 4.4.2 for examples of hedges used in spoken corpora of English). Pragmatic devices are also the focus of another burgeoning subdiscipline, that of variational pragmatics, in which pragmatic variation is investigated in geographical and social space from a synchronic perspective (cf. Schneider and Barron 2008). An example is Plevoets et al.'s (2008) corpus study using data from the 10-million-word Spoken Dutch Corpus (Corpus Gesproken Nederlands) on the distribution of familiar T pronouns and polite V pronouns used as forms of address in contemporary Netherlandic and Belgian Dutch, taking into account the sociolinguistic variables of register, region, age, sex and educational and/or occupational level.

5.4 Limitations of corpus work in sociolinguistics

In spite of the increasing application of corpus linguistic techniques in sociolinguistically motivated studies, McEnery et al. (2006) have drawn attention to three problems besetting such interdisciplinary research.

Quote 5.2 Limitations of corpus work in sociolinguistics

> While sociolinguistics has traditionally been based upon empirical data, the use of standard corpora in this field has been limited. The expansion of corpus work in sociolinguistics appears to have been hampered by three problems: the operationalization of sociolinguistic theories into measurable categories suitable for corpus work, the lack of sociolinguistic metadata encoded in currently available software and the lack of sociolinguistically rigorous sampling in corpus construction.
>
> (McEnery et al. 2006: 108)

5.4.1 Operationalisation of sociolinguistic theories into measurable categories for corpus investigations

This issue is complicated by the fact that the discipline of sociolinguistics, underpinned by a diverse array of approaches and methods, does not have a unified theory: 'Sociolinguistics needs to theorise local social relationships as well as global social structures, the particular moment of social action as well as the dynamics of large-scale social change. The concept of a unified theory is ideologically alien to sociolinguistics, premised as it is on diversity and resistant to hegemony' (Coupland 1998: 113). Moreover, as noted in Section 4.2, corpus linguistic techniques may not be able to capture particular moments of social action as they treat the text as 'snapshots' of the product of interaction rather than as an unfolding discourse process. However, in spite of these limitations, studies in the interactional paradigm show that the identification of routinised speech acts can be gleaned from corpus data and that an ethnographic perspective lends insight into interpretation of the discourse.

The most pertinent and complex issue seems to be the perceived limitations of corpus linguistics to shed light on 'the dynamics of large-scale social change'. De Beaugrande illustrates the drawbacks of corpus linguistics for such sociolinguistic enquiries on the question of multiculturalism with reference to data from the Bank of English.

Quote 5.3 De Beaugrande on corpus data and sociolinguistics

> ... corpus data can help sociolinguistics engage with issues and variations in usage that are less tidy and abstract than phonetics, phonology, and grammar and more proximate to the socially vital issues of the day, especially during the sweeping social change towards the close of the 20th century. Corpus data can help us monitor the ongoing collocational appropriation and contestation of terms that refer to the social conditions themselves and discursively position these in respect to the interests of various social groups.
>
> (*continued*)

> One particularly vital contest is currently between accepting or resisting *multiculturalism* in societies that have traditionally been regarded, albeit implicitly, as monocultural. In the Bank of English data, the term *monocultural* never occurred to designate a society centred on a single culture: the four occurrences concerned the agricultural sense of 'raising a single crop'. A deeper issue arises here, one explored in Critical Discourse Analysis: the mainstream ideology is kept invisible and treated as simply the neutral zero grade or centre from which all differing cultural positions constitute objective deviations (cf. Giroux 1992; Fairclough 1995). The proponents of monoculturalism and opponents of multiculturalism see the value of keeping their own ideologies of greed, selfishness, intolerance, and aggression comfortably outside public discourse.
>
> In contrast, *multiculturalism* frequently occurred in a limited range of variations serving diverse social interests. At least its presence was generally acknowledged, viz.:
>
> 13. more fully portraying the multicultural nature of Britain's society.
>
> Predictably, the socially charged variations of *multiculturalism* were hotly contested between a welcome opportunity … versus a deplorable disruption. I italicise the collocates that might signal discursive interests:
>
> 16. contribute to our *strength* as a multicultural society *that welcomes diversity*
> 20. *quality is sacrificed* for multicultural *equality*
>
> (de Beaugrande 2002: 135–6)

De Beaugrande's key observation is that what is left unsaid in corpus data obscures the true picture in CDA studies attempting to get to grips with 'the dynamics of large-scale social change'. This 'silence' would have to be made visible by other means. In order to analyse such sweeping social changes as multiculturalism, it would be necessary to adopt the type of multidimensional analysis proposed by Baker et al. (2008) embracing social, political, historical and cultural contexts of the data going beyond the purely linguistic elements of diachronic corpora (see Quote 4.6).

At the level of local social relationships Sealey (2009) has brought together insights from realist social theory, complexity theory and corpus linguistics to focus on one aspect of identity, that of self-representation, in oral history interviews. Realist social theorists seek to understand rather than observe directly the generative mechanisms responsible for differential social behaviour. Sealey, moreover, views the interviews as representing 'examples of

discourse as a complex system' entailing 'a much more complex and diverse set of language-using patterns than the 'core grammar' of formal approaches' (Larsen-Freeman and Cameron 2008: 99).

The corpus comprised 144 transcribed interviews by two oral historians of narrative accounts, including questions on childhood memories, family life, etc., of a diverse range of residents of Birmingham recorded at the turn of the millennium. Unlike in quantitative, variationist analyses, Sealey's aim is not to see which demographic variables correlate with particular linguistic features, but rather to examine interaction between the variables mapped against dynamic, systems-based realities with particular reference to the central role human agency plays in the choices interviewees' make about how to formulate an account of their life experience.

Example 5.6 The realist approach exemplified

Sealey uses the category of gender to exemplify this multi-method approach. A case-oriented methodology is adopted which allows for a holistic comparison of cases according to different combinations. First, the demographic metadata were entered into a table, with each row representing a case. In the next stage, keywords identified using WordSmith were examined. For example, *lovely* was used four times as often by the female speakers (441 occurrences) as by the men (110 occurrences).

Sealey challenges the kinds of categories with homogenising limitations which might interpret such findings as the use of *lovely* indexing femininity by reference to complex causation and looking at counter-examples. From a case-based perspective, the two men who used *lovely* portrayed similarities across several dimensions; both were born in Birmingham in the 1920s, neither continued their education beyond secondary school, etc. Sealey also used the *WMatrix* tool to examine the semantic domains in the interviews in which *lovely* was not used by females.

Sealey's exemplification of this multi-method approach is thus a first step 'to combine corpus analysis and realist-derived sociological analysis through software adapted to accommodate both kinds of data' (p. 227).

5.4.2 Encoding of sociolinguistic data using current software

Although the collection of metadata on sociolinguistic variables still remains a very time-consuming and labour-intensive task, one area in which the field is advancing is in the adaptation of more sophisticated software for dealing with sociolinguistic metadata (see Kretzschmar et al. 2006 for a detailed overview of a corpus management system for encoding and refining searches by metadata).

However, the question that now arises is how far one should go in the markup of the corpus with such variables. As Biber et al. (1998) have remarked, for finely tuned analyses, it would be necessary not only to consider which utterances are spoken by males and females, but also whether the person being addressed is male or female since research has shown that this is an important factor in how a female or male speaks to a person. Going a stage further, Goffman's (1981) notion of ratified and unratified hearers, i.e. anyone who is intentionally or unintentionally party to the discourse, might well have a bearing on the analysis and therefore be another variable to consider.

5.4.3 Sociolinguistic sampling procedures in corpus compilation

Meyer (2002) has noted the importance of probability sampling methodology used for selecting speakers to ensure that the number and type of people are representative of the population as a whole.

Concept 5.6 Probability sampling methodology for BNC

> Crowdy (1993) states that for selecting speakers for the BNC Great Britain was divided into 12 dialect regions. Within these regions, 30 sampling locations were selected at which recordings were made based on the speakers' ACORN profile (A Classification of Regional Neighbourhoods), providing demographic information. Compilers of the BNC selected speakers of various social classes based on this profile and also controlled for other variables such as age and gender.
>
> (Adapted from Meyer 2002)

However, Meyer (2002) notes that for creation of the BNC only 124 speakers of varying social classes were selected from these 12 dialect regions, thus underscoring the difficulty of creating nationwide corpora balanced according to sociolinguistic variables such as region and social class. Meyer therefore advocates using smaller, more regionally focused corpora for studying dialectal social and regional variation. Nevertheless, as Aston and Burnard (1998) point out, the BNC is perfectly adequate for comparisons to be made across broad social groups such as 'North', 'Midlands' and 'South'.

Another concern raised relates to the sampling methodology employed for corpora of non-standard varieties of English. Schmied (1990) notes the potential mismatch between the stylistic sampling categories based on text types used for compiling corpora derived in native communities (ENL corpora) and the sociolinguistic ones based on speaker/writer identity for corpus compilation of non-native communities (ESL corpora), in this particular case the ICE Corpus of East African English. With reference to the Scottish Corpus of Texts and Speech

(www.scottishcorpus.ac.uk), which, in fact, is a blanket term for several varieties ranging from the more middle-class *Scottish Standard English* to working-class *Scots* at the other end of a continuum, Douglas (2003) cautions against sampling on the basis of stratified genres, as is usually advocated. As *Scots* only tended to occur in letters pages and certain types of feature articles in newspapers, Douglas notes that sampling solely on the basis of newspaper genre would be misleading and that it is necessary to have 'inside knowledge' of non-standard varieties of English to devise appropriate sampling methodologies.

5.5 Conclusion

In spite of perceived limitations, what the studies above demonstrate is that more sociolinguistic-sensitive approaches are filtering through into corpus-based analyses, and it is encouraging to note that this also applies to languages other than English. Researchers are now beginning to make use of a sophisticated 'toolkit' of sociolinguistic approaches for analysing corpus data, combining conversation analysis and interactional sociolinguistics with fine-grained linguistic analyses, the type of eclectic approach advocated by Sarangi and Roberts (1999) and others for analysing institutional interaction. Such data could also be complemented by the use of multimodal analysis discussed in Section 4.5 for incorporating physical, visual and kinetic features into corpus-based sociolinguistic analyses.

While such corpora provide opportunities to explore hitherto uncharted corpora domains they also bring new challenges in the form of adaptations of present corpus-building techniques and search queries to accommodate new non-standard corpora (McEnery and Ostler 2000).

Further reading

Baker, P. (2010) *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press. This volume presents a very clear overview of corpus linguistic applications to sociolinguistics for those who are new to the field.

Rühlemann, C. (2007) *Conversation in Context: a Corpus-Driven Approach*. Amsterdam: John Benjamins. This monograph presents a contextual analysis of speech data, specifically conversation, from the British National Corpus. Integrating corpus linguistics, discourse analysis and sociolinguistics, Rühlemann uses a situational framework for the analysis.

Yates, S. (2001) Researching internet interaction: sociolinguistics and corpus analysis. In M. Wetherell, S. Taylor and S. Yates (eds) *Discourse as Data. A Guide for Analysis*, pp. 93–146. Milton Keynes: The Open University. This chapter presents a step-by-step breakdown on using corpus linguistic techniques to analyse the construction of identities in CMC interaction.

Part III
Applications of Corpora in Research and Teaching Arenas

6
Applying Corpus Linguistics in Research Arenas

This chapter will:

- Examine the application of corpus linguistics in key research arenas (ELF, professional communication, forensic linguistics, corpus stylistics, translation, learner corpora and SLA, lexicography and testing)
- Discuss the main approaches adopted for corpus analysis in the different arenas
- Survey key issues related to different philosophical and theoretical underpinnings
- Map out some pathways for future research

Sections 6.1–6.6 take an *applied* corpus linguistic perspective to various research areas. Chapter 7 takes an *appliable* perspective, a term I appropriate from Halliday (2008) to refer to the applications in using corpora for various kinds of teaching purposes. Section 6.7 on corpora for lexicographic purposes and Section 6.8 on corpora for testing purposes provide a bridge between the research-oriented focus of Chapter 6 and the pedagogic-oriented focus of Chapter 7, straddling the *applied* and *appliable* corpus linguistic interface.

6.1 English as a lingua franca (ELF) research

The previous chapter has discussed corpus studies on a number of varieties of inner and outer circle world Englishes, situated within the variationist paradigm (see Section 5.3.2). ELF has not been included in the preceding chapter as its status as a 'variety' is open to serious question, an issue discussed later in this section.

6.1.1 Definition and status of ELF

Concept 6.1 English as a lingua franca (ELF)

English when defined as a lingua franca serves as a means of communication among speakers in the 'expanding circle' of world Englishes, i.e. in countries such as Spain, Sweden, China, etc. where English has no official status, but is used as a lingua franca among non-native speakers. Moreover, it should be noted that ELF cuts across outer/expanding circle distinctions since speakers from outer circle countries, e.g. India, who use English intranationally also participate in global uses of English (Kachru 1986).

Nowadays there are more non-native speakers of English than there are native speakers; common estimates are that 400 million people speak English as a first language, with another 300–500 million speaking English as a fluent second language and perhaps 750 million as a foreign language. Beneke (1991), cited in Seidlhofer (2004), estimates that about 80 per cent of verbal exchanges in which English is used as a second or foreign language do not involve any native speakers of English.

Given the increasing status and use of English as an international language, Seidlhofer (2001a) and Mauranen (2007) make a strong case for compiling ELF corpora to investigate the language used among non-native speakers in order to offset the over-reliance on native speaker corpora, which, according to Seidlhofer (2004: 214), has tended to devalue the language of non-native speakers and hence perpetuate the notion of linguistic imperialism, giving 'deference to hegemonic native-speaker norms'. To redress this imbalance, and also complement the ICE parallel subcorpora representing inner and outer circle varieties of English, it would seem that ELF corpora would be a fruitful avenue for exploration. However, whether ELF constitutes a variety of English has been hotly debated.

Concept 6.2 Is ELF a variety of English?

Although Mauranen (2003b) describes ELF as a variety, O'Keeffe et al. (2007: 29) point out that ELF interactions may involve a Chinese speaker interacting in English with a Korean speaker, or a Danish speaker interacting in English with a Dutch speaker and that 'we may need to describe "many ELFs" to get anywhere near an accurate picture of the global uses of English'. Likewise, Mukherjee (2005), cited in Hundt (2009), does not see ELF as a variety, but rather as a conglomerate of variants and also as a kind of interlanguage 'including all kinds of variants that we find in different learners with different L1 backgrounds and at various competence levels' (Hundt 2009: 454).

Granger (2009: 25) echoes both of the above views, commenting on varieties of ELF and the close relationship between EFL and interlanguage:

> Research on learner corpus data, which are very close to ELF data, points to a very high degree of L1-specificness, especially as regards lexis and lexico-grammar. Like Mackenzie (2003) I therefore think that a description of ELF features is unlikely to reveal ELF; a Romance ELF will probably turn out to be quite different from a Germanic ELF, which in turn will be quite different from an Asian or South African ELF.

6.1.2 Corpus projects on ELF

Two major, yet different, initiatives are underway to compile ELF corpora. The Vienna-Oxford International Corpus of English, VOICE, is targeted to contain 1 million words of spoken ELF from various professional, educational and informal settings (http://www.univie.ac.at/voice). Another major endeavour in this field is a more specialised corpus project, the Corpus of English as a Lingua Franca in Academic Settings (ELFA) at the University of Helsinki (http://uta.fi/laitokset/kieket/engf/research/elfa/). (See Mauranen et al. (2010) for a description of the rationale and design of the ELFA corpus.) The purpose for focusing on spoken language rather than written in these corpus projects is that written text may not be a true reflection of ELF due to various editing processes.

Corpus linguists working in this area (cf. Seidlhofer 2001a; Mauranen 2007) consider ELF as a versatile, sophisticated tool for successful intercultural communication (rather than as a pidgin language with a restricted vocabulary and style), among speakers who 'can express their identities and be themselves in L2 contexts without being marginalised on account of features like foreign accents, lack of idiom, or culture-specific communicative styles as long as they can negotiate and manage communicative situations successfully and fluently' (Mauranen 2003b: 517). However, see Section 6.2.1 for an example illustrating how difficult it could be to gauge what constitutes successful intercultural communication.

Although the ELFA corpus is quite specialised, which may make it easier to identify specific features of successful ELF interaction in an academic setting, the VOICE corpus is far more wide-ranging in its scope in terms of speech events, interlocutors and functional settings. This means that it is likely that there may be 'many ELFs' and that ELF in its broadest sense is not amenable to codification by means of a general set of parameters such as those that exist for establishing standard varieties of English (see Concept 6.2). Some considerations that Seidlhofer has suggested for pinning down characteristic features for ELF are given below, but these are just the initial steps in defining this

burgeoning new type of English, which have enormous implications for the teaching of English for global communication (see Section 7.4.5).

Quote 6.1 Seidlhofer on common features for ELF

- What seem to be the most relied-upon and successfully employed grammatical constructions and lexical choices?
- Are there aspects which contribute especially to smooth communication?
- What are the factors which tend to lead to 'ripples' on the pragmatic surface, misunderstandings or even communication breakdown?
- Is the degree of approximation to a variety of L1 English always proportional to communicative success, or are there commonly used constructions, lexical items and sound patterns which are ungrammatical in standard L1 English but generally unproblematic in ELF communication?
- If so, can hypotheses be set up and tested concerning simplifications of L1 English which could constitute systematic features of L1?

(Seidlhofer 2001b: 78–9)

A linguistic analysis of ELF corpora can be enhanced through taking into account other factors. For example, Bjørge (2010) investigated both verbal and non-verbal aspects of backchannelling in a corpus of 13 video recordings of ELF simulated negotiations (see Section 4.5.2 for another multimodal analysis of backchannels, and also Section 5.3.2 for an analysis of backchannels comparing outer circle varieties of English). Meanwhile, Mauranen et al. (2010) bring a more ethnographic perspective to their ELF data in the SELF (Studying English as Lingua Franca) project, which seeks to explore discourse phenomena such as negotiation of meanings and sources of misunderstanding.

Quote 6.2 ELF corpus data explored from an ethnographic perspective

The SELF project seeks to relate a linguistic perspective to a more social one, looking for interactive and adaptive processes in action, and talking to the participants about them. Data for SELF is collected from interrelated speech events, interviews and observation as well as written documents. The micro-analytic perspective enables researchers to explore ELF as situated language use, in this way connecting the global phenomenon of ELF to local practices of adaptation and assimilation – even resistance.

(Mauranen et al. 2010: 187)

6.1.3 Regional ELFs in intercultural communication

However, many questions on the very nature and development of ELF remain to be answered (see Concept 6.2). The most contentious one is whether ELF will emerge as a distinct variety of English in its own right, as a kind of 'supranational standard' (Chambers 2000: 285, cited in Seidlhofer 2004). In fact, according to some linguists, this already seems to be taking place at a regional level with the emergence of a 'Euro-English' as a lingua franca in the European Union (EU) (Graddol 2006), with official EU documentation sanctioning lexis such as 'concertation' for 'concerted actions', resulting in an endonormative model of lingua franca English, whose own usage rather than that of standard UK English drives the norms for acceptability. But there are also critical voices who dispute the existence of a 'Euro-English'. One such linguist is Mollin (2006: 1), who views it as a 'Yeti of English varieties: everyone has heard of it, but no one has ever seen it', reaching this conclusion based on her study of a 400,000-word corpus of interactions among users of English in the European Union. However, Seidlhofer (2010) presents a rebuttal to Mollin's findings based on the grounds of her conceptual framework for corpus design and focus on formal rather than functional features in the analysis.

Quote 6.3 Is there a Euro-English?

> Speakers stick to native-speaker standard usage and make individual 'errors', if one wishes to name them so, depending on mother tongue and English competence generally. There were hardly any common features that united lingua franca speakers, even in a context such as the EU, where speakers use English frequently, interact with each other and do not have the opportunity to negotiate a common standard.
>
> (Mollin 2007: 48)

Another region where ELF is deemed to be taking hold is in China where a 'China English' is considered to be emerging in step with the country's increasing economic importance on the world stage. Kirkpatrick and Xu (2002) note that of the 350 million Chinese currently learning English, the great majority are likely to use it with other non-native speakers. They envision a 'China English' with Chinese discourse and rhetorical norms inevitably developing (in fact, they suggest that there may also be another variety of English developing alongside it to accommodate to native speakers of English).

In sum, there is without doubt still much work to be done in identifying and codifying salient features of ELF which, as Seidlhofer and Mauranen have pointed out, not only concern phonology, lexis and syntax, but also pragmatic strategies in intercultural communication, for which non-visual

and ethnographic data are now being collected. The few corpus-based studies conducted so far indicate that while it may not be possible to isolate features of a pan-global ELF, there may be new varieties emerging such as 'China English' or 'Euro-English' for which some codification has been undertaken. See Section 7.4.5 for discussion of ELF from a pedagogic perspective.

6.2 Research in business and health care contexts

As far as modes of communication are concerned, corpus studies are at present dominated by those of spoken business English, with a few studies examining interactions between health professionals and patients.

6.2.1 Interactions in the business context

This section first takes up the theme of intercultural communication with its implications for ELF, followed by a discussion on transactional vs relational talk and the importance of the ethnographic dimension in researching business contexts.

Intercultural communication

Seidlhofer (2004) has argued that studies on ELF seem to show that there is a greater tolerance for pragmatic infelicities on the part of ELF speakers, whose main use of ELF is instrumental in order to get their message across. The two studies discussed below will be used to examine this point in more detail.

Bowles's (2006) study of service calls to bookshops to negotiate requests examines NS-NNS telephone calls, an interaction that Bowles states NNS might have particular difficulty with 'in the absence of paralinguistic cues that are characteristic of face-to-face interaction' (p. 333). Bowles's corpus consists of 40 calls made by English native speakers to English bookshops and of 50 calls made by Italian native speakers in English to English bookshops. Like Koester, he also draws on conversation analysis, in this study utilising the turn construction unit (TCU) from CA, a unit of talk which is complete semantically and syntactically, to investigate the reason-for-call, a particularly difficult turn for NNS to negotiate with correct management of pre-sequences.

Example 6.1 Comparison of TCUs by NS and NNS

Bowles' analysis reveals that both TCU pre-requests and somewhat extensive stories were used more frequently by Italian callers than the NS callers. Bowles suggests that the Italian callers are transferring pragmatic conventions from Italian to these service encounters, thus highlighting how useful quantitative corpus studies can be for uncovering recurring pragmatic features providing insights into intercultural communication.

Call 9 (NNS corpus)

1. R: xxxxx books good afternoon
2. C: .h oh hi ehm: **i wonder if you can help me** ← **TCU pre-request**
3. C: **mch i would like to: :make a present eh: :**
   (.) ← **Story**
   for a friend of mine. hh a: :nd m: :i:s really
   fond of <u>cats</u> .h and I was wondering if you ← **TCU pre-request**
   have anything like: :cat books like with. H
   a lot of pictures and maybe some: : .hh ah: :
   informa[tion about the races e: an all that (.)
   you know
4. R: (so) something about breeds on cats like a photographic book on breeds

(Bowles 2006: 346)

Opening sequences have also been examined in the Hong Kong Corpus of Spoken Business English, HKCSE, which has been prosodically transcribed (see Section 4.4.1). Cheng (2004) comments on instances of pragmatic failure and also lexico-grammatical problems in some utterances on the topic of checking out of a hotel.

Example 6.2 Pragmatic infelicities in the HKCSE

In B004, the front office staff's question 'Mister T_ (pause) have you get the minibar key' (line 3), which is lacking a *please* and hedging, coupled with the fact that it is the first utterance of the interchange, makes the speaker sound rather abrupt towards this guest. The question also contains a grammatical mistake ('have … get').

B004
3b:// ↗ HAVE you get the minibar KEY //
4b:// ↘ I wasn't GIven one //

(Cheng 2004: 146–8)

Bargiela-Chiappini et al. (2007: 94) have called for 'a large scale corpus-based investigation of the language used in BELF (Business English as a Lingua Franca) interactions', stating that 'it would help to pinpoint those areas where native and non-native varieties of business English are different (and therefore potentially problematic)'.

While the example above from Cheng involving a transactional exchange seems to be fairly uncontroversial, the opening sequences of Italian callers identified by Bowles would appear less clear cut to classify as instances of pragmatic failure. Although the rather long-winded opening with a personal story is breaking Grice's Maxim of Quality (Do not make your contribution more informative than required), it may well be succeeding on the phatic level of communication involving relational talk. These two examples illustrate the types of issues that need to be discussed, and how difficult it may be to define what constitutes successful communication in ELF intercultural exchanges. A way forward would be to examine pragmatic infelicities in ELF from an interactional perspective within the four broad parameters of phatic exchanges, transactional talk, relational talk or transactional-cum-relational talk, to determine to what extent ELF is acceptable (see the following section).

Transactional, relational and institutional talk

Koester's (2006) corpus-based research on workplace discourse, much of it drawn from CANBEC, has already been discussed in Section 5.2, with reference to analyses illustrating its anchoring in conversation analysis and ethnography (see also Handford 2010b; McCarthy and Handford 2004). Another analytical framework Koester draws upon is the turn-by-turn development of talk relating to transactional and relational goals. See Section 3.4.1 for a similar framework drawn up for the CANCODE data on which CANBEC is based.

Concept 6.3 Koester on transactional vs relational talk

Koester (2006: 33–4) divides transactional talk into unidirectional and collaborative discourse. Unidirectional discourse covers briefings, service encounters, procedural and directive discourse, requesting and reporting. Collaborative discourse includes arrangements, decision-making, discussing and evaluating and liminal talk.

Relational discourse, which can be found at the level of an entire conversation or individual words, is classified as follows (pp. 55–6):

1. Non-transactional conversations: office gossip and small talk
2. Phatic communication: small talk at the beginning and end of transactional encounters
3. Relational episodes: small talk or office gossip occurring during the performance of a transactional task
4. Relational sequences and turns: non-obligatory task-related talk with a relational focus
5. Interpersonal markers: modal items, vague language, hedges and intensifiers, idioms and metaphors

As Koester (2010) shows, corpus linguistic techniques are ideal for uncovering quantitative data such as keywords, collocations, chunks, etc., which can then be explored from a more qualitative perspective. For example, the fundamental basis of workplace communication as essentially a transactional, goal-oriented activity has been borne out by corpus-based studies on CANBEC and Nelson's (2006) 1-million-word Business English Corpus, BEC.

Example 6.3 Goal orientation identified through corpus linguistic techniques

Koester (2010: 45–69) summarises corpus linguistic findings from both CANBEC and BEC to show that business language is essentially goal-oriented action:

- the dynamic and action-orientated keywords in BEC, e.g. 'manage', 'operate'
- the unusually high frequency of the personal pronoun 'we' with entities (products and companies) being talked about more than people in CANBEC and BEC
- the use of deontic modals in highly frequent chunks, e.g. 'we need'/'we need to'/'we need to do' to talk of necessary or desirable actions in CANBEC and BEC

At the same time, workplace discourse is very much relational, displaying many of the interpersonal markers such as personal deixis, modals, indirectness and hedging outlined in Concept 6.3. However, interpersonal markers such as idioms and metaphors are not always relational talk marking convergence between interlocutors (see Concept 6.3, point 5). Handford and Koester's (2010) research shows that they can also occur in conflictual business meetings, marking divergence. Many also signal intensity, e.g. the metaphorical phrase *the only thing that fouled up with this letter*, performs the function of negative evaluation but, simultaneously, the basic meaning of the metaphor (i.e. 'dirty' or 'unpleasant') indicates the strength of feeling (p. 42).

Example 6.4 Relational talk in CANBEC

The qualitative analysis reported in McCarthy and Handford (2004) reveals speculative and hypothetical uses of *may* and *might*, which they argue are 'important parts of the collaborative and convergent enterprise of consensus-making' and also 'face-protecting both for those who speculate and

(*continued*)

those who respond' (p. 182). *May* is found to occur in combination with other hypothetical expressions, vague language and hedges:

Extract 7
[Meeting between the sales staff of an IT company and a potential client. The latter is the Managing Director of an Internet Sales company]
<$1> I *guess* you'll have to speak to Bob and and and to to James and and and *kind of* look at what you think you *may* have coming up.
<$2> Yeah.
<$1> And then we can get together again and actually you know finalise *something* and and and move forward.
<$2> Okay.
<$1> Yeah.
<$2> Yeah.
<$1> So it *may well be* that it's it's it makes financial sense to go with a rack and a half and put all of your existing servers into a rack. It *may* be better that erm you keep those ones and *maybe* just a half rack for the future. Or it *may* be *a bit better* just to keep buying individual collocations *as and when* you need them.
<$2> Right okay.

(McCarthy and Handford 2004: 182–3)

One point of interest is that while McCarthy and Handford found business English to share many relational features with conversation in terms of its 'orientation towards comity, convergence, and satisfactory and non-threatening relationships' (p. 187), the keywords analysis (with words such as *okay* and *problem* found to be key), shows it to be different from everyday conversation and that it remains essentially an 'institutional form of talk'. The prevalence of problem-solving words, e.g. *make a decision*, has also been found in other sites of business communication lending weight to the concept of 'institutional talk' (see Case Study 8.6 on discursive practices in the construction industry).

Ethnographic dimension

Given the context-sensitive nature of business discourse, corpus data such as CANBEC are most fruitfully analysed with respect to ethnographic considerations such as in the case of Handford's (2010b) research on business meetings from CANBEC with attention to the social, professional and discursive practices of this genre (see Bhatia 2008). Handford conducted follow-up interviews to tease apart discursive practices signifying 'recurrent patterns of linguistic behaviour that are decipherable in transcripts of business meetings' (p. 66)

and, on the other hand, strategies which are envisioned as less normative than practices and constitute somewhat atypical language to achieve a particular communicative intent (see Section 1.3.2 on difficulties with interpretation).

Example 6.5 Handford on social, professional and discursive practices

… an example of a professional practice is setting up the supply chain process and a social practice in business could be 'being a manager'. As such, social practices, and to a lesser extent professional practices, are more difficult to locate in meeting transcripts when compared to discursive practices. Nevertheless, overall the topic of a meeting may explicitly reflect the professional practice, whereas the discursive practice often relates to particular stages and actions within the meeting itself (p. 67). (See Figure 6.1.)

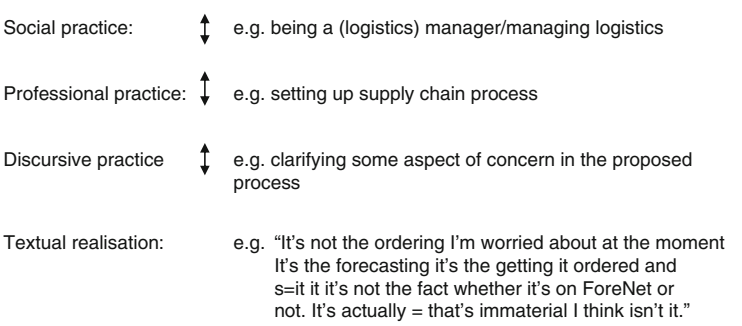

Figure 6.1 The relationship between practices, text and context (Handford 2010b: 67)

The Wellington Language in the Workplace Project, LWP (Stubbe 2001) also adopts an ethnographic approach to data collection methods and relationships with the participating companies. The sociolinguistically motivated analyses of this corpus have examined transactional and relational talk (Holmes 2004, 2005) and the interpersonal device of humour (Marra and Holmes 2002).

Quote 6.4 Data collection for the Wellington LWP

The central principle of the method we have developed is to use a participatory approach which entails the active involvement of the volunteers (key informants) from a particular team or section of an organization. Essentially, this has meant establishing and maintaining an ongoing dialogue with the individuals and organisations involved, based on the principles of action

(*continued*)

research and appreciative enquiry, and giving both our key informants and other participants ownership over the collection and subsequent uses of the data. This approach has made it possible for the research team and participants to collaborate in setting the research agenda and exchanging relevant information, and has provided a sound basis for the people and organisations concerned to reflect on and develop their own communicative practices. The basic data collection model moves through four distinct stages:

(i) making contact/establishing the research relationship;
(ii) collecting ethnographic information and recording talk;
(iii) initial processing and analysis of the data; and
(iv) providing reflexive feedback.

(Stubbe 2001: 5)

6.2.2 Interactions in the health care context

Transactional, relational and institutional talk

Building on the seminal work of doctor–patient interviews by Candlin, Bruton and Leather in the 1970s, several corpus studies exemplify how what may appear to be unidirectional transactional talk can be skilfully transformed into more seemingly collaborative talk through the blending in of relational features. The main speech act performed by the interlocutor in the advisory role would usually be some form of directive. The corpus studies outlined below illustrate the ingenious ways in which health care professionals exploit language to disguise directives.

A corpus study by Skelton et al. (2002) on how doctors and patients use first person pronouns in primary care consultations revealed that doctors in the study used 'we' in an inclusive way, e.g. *I think we should actually change your tablets*. While such language can be interpreted as promoting a patient-centred atmosphere, doctors also frequently used 'we'll' to initiate a discussion of action giving rise to ambiguity as to whether inclusive or exclusive 'we' was intended. Of note, is that this type of phraseology was first uncovered by Candlin et al.'s (1981) analyses of authentic doctor–patient interaction for designing communicative language teaching materials. They noted alternative utterances for realising the same function, asking how these might affect the doctor–patient relationship, and whether it should be:

'Nurse, give this patient anti-tetanus, would you.'
or
'I think we'd better give you a little jab, Mr. Smith, just to be on the safe side.' (Candlin et al. 1981: 106)

Other types of directives have been examined by Adolphs et al. (2004) in a corpus of institutional telephone conversations between callers and advisers in the UK's 'NHS Direct' health advisory service. This is a small corpus of 61,981 words (35,014 constituting the health professionals' corpus and 26,967 words for the patients' corpus). Adolphs et al. (ibid.: 13) state that as the corpus is 'relatively specialised and coherent' it serves its purpose as a 'preliminary vignette' into NHS Direct interactions (see Section 1.2.1 for discussion on size of a corpus). The analysis borrows in a general way from CA as it looks at how the clinical encounters are jointly constructed by the interactants and how language is used to facilitate the practical procedures and goals of clinical work, mainly achieved via strategies of politeness and the language of convergence. Closings, i.e. 'convergence codas' containing a summary of the preceding conversation, were commonly found at the end of the call with the aim of encouraging the patient to adopt a course of action. Politeness markers were signalled by backchannel responses, 'may' (functioning as both a politeness device and epistemic softener, e.g. …*stopping it tonight may not reduce your symptoms tonight*) and 'if' with a similar function to 'may'. 'By introducing hypotheticality into the discourse it creates options for the patient and it also softens or mitigates any advice that is given' (p. 18).

'If' conditionals are also the subject of another corpus-based study of medical discourse across three different genres: research articles, journal editorials and doctor–patient consultations (Ferguson 2001). In the doctor–patient consultations out of a total of 77 conditionals, polite directives of the pattern present + future were found to be the most frequent with 20 occurrences, e.g. *If you go outside, Sister will fix things up*. Interestingly, 10 of these were without apodosis, e.g. *Perhaps if you tell me about these spots*.

The above studies demonstrate there to be some commonalities between interactions in the business and health care contexts, with their respective foci on collaborative convergence strategies for giving advice and directives, which could be viewed as essentially relational talk masquerading as institutional talk.

Vague language, identified as a feature of phatic communication in service encounters (McCarthy 2000) and offices (Koester 2007), also surfaces, somewhat surprisingly, as a regular feature in health care contexts but is not merely of a phatic nature. Adolphs et al. (2007) find this kind of institutional talk to have two important functions in nurse–patient communication in the same corpus of NHS Direct referred to earlier (see also Harvey and Adolphs 2011).

Concept 6.4 Functions of vague language in medical interactions

Like hypotheticality, vague language, especially the phrase 'or anything', on the part of the nurse opens up a discursive space for the patient to add

(*continued*)

their own description of the situation. The examples below from Adolphs et al. (2007: 66–9) illustrate the common discourse strategies nurses use for diagnostic purposes.

NHS Nurse: Er any intense headache or mental confusion or anything?
NHS Nurse: No shortness of breath or gasping for breath or anything
NHS Nurse: And so there's no swelling anywhere to your face or anything?

The other function of vague language is to downplay any potentially distressing serious diagnoses:

NHS Nurse: Cos we have to kind of un….We always do like the worst case scenario and
and work downwards.
Patient: All right (laughs) Okay.

The same kind of phenomenon has been noted among general practitioners by Skelton and Hobbs (1999: 109). They found that doctors often used the softening phrase 'what they call', sometimes coupled with mitigating phrases to diminish threats and reassure patients: e.g.

> Might have a little of *what they call* erosion on the neck of the womb
> Happens is you get a bit of *what they call* intestinal hurry anyway so the food

Candlin (2006) has called into question a situation of dominance of the nurse over the patient, and the above corpus findings would corroborate the view that the patient's contribution to the discourse is a voice to be heard, an aspect that is taken up in more detail in the following subsection.

Ethnographic dimension

Harvey and Adolphs (2011) have underscored the importance of integrating both qualitative and quantitative approaches to data analysis in health care communication. One early study which used multiple methods for studying the effectiveness of communication between patients suffering from cancer and a range of health service carers was that by Thomas and Wilson (1996). The ethnographic, qualitative research took three forms: a preliminary study of the cancer services in two district health authorities, long-term case studies of the interaction between cancer patients and their carers, and research interviews. A 2-million-word corpus was compiled from the out-patient clinic sessions involving patients, their relatives and a consultant. The quantitative research took the form of 270 questionnaires administered to patients in the two health authorities to find out, for example, how much opportunity

patients had had to talk about their feelings concerning their illness, and their perception of the degree of emotional support offered. Through a comparison of questionnaire responses with corpus frequency counts of various semantic categories such as pronouns, hedges, boosters and downtoners, Thomas and Wilson were able to provide evidence that the language produced by Dr A was interactive, interpersonally oriented and informal, whereas Dr B's was more disease-centred and technical, findings which were in alignment with other aspects of the ethnographic data.

6.2.3 Implications and future directions

Given that professional communication, in the business world especially, is highly context and situation specific, one question that arises regarding the corpus findings of a particular professional genre is how applicable these would be across different settings in view of the different social, professional and discursive practices that may exist not only across individual companies but also across different cultures. Several subcorpora of business meetings now exist: Bargiela-Chiappini and Harris's (1997) corpus of business meetings based on approximately 18 hours of business meetings recorded in Great Britain and Italy; Nelson's (2006) Business English Corpus containing recordings of business meetings conducted in the UK and USA; the HKCSE containing a subcorpus of meetings collected from companies in Hong Kong, and CANBEC based on recordings obtained from UK companies, but their findings may well reveal different discursive practices. One fruitful avenue for future exploration would be more ethnographically motivated investigations of corpora, such as those of Handford (2010b) and Holmes (2004, 2005), of the same genre across different cultures to provide insights into the field of intercultural communication.

As this chapter has shown, the emphasis of corpus research has been on spoken communication of various kinds, particularly relating to the field of business. What are needed are more corpora of a wide variety of professional written documentation. One such initiative is the ongoing compilation of a 100-million-word Corpus of Professional English in science, engineering, technology and other fields under the auspices of the Professional English Research Consortium, PERC (Noguchi et al. 2006). Another professional corpus which stands at nearly 4 million words is the Hong Kong Financial Services Corpus, including subcorpora of annual reports, interim reports, codes of practice and rules (cf. Cheng 2009). Research on other modes of professional communication such as CMC are now being set up (see Atkins and Harvey's (2010) project on an e-mail corpus of messages from young people sent to GPs in confidence, which is discussed in further detail in Case Study 8.7).

Although corpus linguistics has not made as many inroads into the field of professional communication as it has in general ELT and both EGAP and ESAP (see Section 7.1), it is expected that in future new corpora of under-researched

genres in under-researched disciplines will result in more corpora not only *of* professional communication, but also *for* professional communication training purposes in the business and health care contexts. One point that should be stressed is that corpora can be an invaluable resource for professional training as corpus findings may have important implications for outcomes. Not only does this apply to the contexts discussed in this section but also to the professional sphere discussed in the following section – forensic linguistics.

6.3 Forensic linguistics research

Corpora have been used in forensic linguistics in the following two main ways: for the attribution of authorship in texts such as witness statements and for the analysis of courtroom discourses, usually with the use of a reference corpus for comparative purposes (corpora have also been used for detecting plagiarism, which can arise in legal cases; see Coulthard 2004; Hunston 2002 on this issue).

Concept 6.5 Forensic linguistics

> Forensic linguistics involves the application of scientific knowledge to language in the context of criminal and civil law. Forensic linguists have an interest in understanding the language of the written law, its complexity and its origin, as well as the use of language in forensic procedures. They also study the judicial process from point of arrest, and through the interview, charge, trial and sentencing stages. For example, linguists are interested in the language of police interviews with witnesses and suspects, and in the language of lawyers and witnesses in cross-examination.
> (Arts and Humanities Research Council: Forensic linguistics, sector overview)

6.3.1 Corpora for attribution of authorship

The lion's share of the work in English in this area involving analysis of register and genre variation has been carried out by Coulthard (1993, 1994, 1995, 2004), director of the Centre for Forensic Linguistics at Aston University in the UK. Coulthard (1993: 86) states that in cases involving attribution of authorship it is usually the defence who has asked the linguist 'to support a claim that an accused is not the (sole) author of a document'. A request by the prosecution 'to provide support for the more difficult claim that a defendant *is* the author' is much more rare. This is because the nature of the corpus evidence that the linguist submits in court is almost always probabilistic. Such evidence is 'currently more usable by the defence, where the need is to show "reasonable doubt", than by the prosecution, where the need is to show "proof"' (p. 87).

One of the most oft-cited cases in which Coulthard's forensic linguistic skills have been successfully applied is in the acquittal of Derek Bentley who was hanged for murder in 1953, but acquitted posthumously in 1999. In McEnery et al.'s (2006: 116) view forensic linguistics is 'perhaps the most applied and exciting area where corpus linguistics has started to play a role because court verdicts can very clearly affect people's lives', and the notorious Bentley case is a prime example of this, as illustrated below.

Example 6.6 Corpus-assisted analysis of register variation

Coulthard (1995) examined the statement attributed to Bentley by the police, but which Bentley claimed to have been heavily influenced by them. Coulthard (1996: 166) is thus questioning the validity of the 'verbatim' record, i.e. the *locutionary* record in Austin's terms. Coulthard (1995) compared Bentley's statement with three other sources (a) statements made by other lay witnesses (b) statements made by police officers (c) a mixed corpus of 1.5 million words of spoken English.

Coulthard and Johnson (2007) report that a marked feature of Bentley's confession was the frequent use of 'then' in its temporal meaning with 10 occurrences in 582 words: e.g.

Chris **then** jumped over and I followed.
Chris **then** climbed up the drainpipe to the roof and I followed.

In contrast, in the witness statements there was only one occurrence of 'then' in 930 words. Strikingly, though, 'then' was found to occur 29 times in the police officers' statements, i.e. once every 78 words on average. Coulthard thus concluded that 'Bentley's usage was at the very least untypical, and thus a potential intrusion of a feature of policeman register, which is related to a professional concern with the accurate recording of temporal sequence' (p. 88). However, Coulthard also consulted a 1.5-million-word reference corpus of spoken English to check the representativeness of the 'ordinary witness' data. That 'then' only occurred 3164 times in the whole corpus, i.e. once every 500 words 'supported the representativeness of the witness data and the claimed specialness of the police and Bentley data' (p. 88).

Coulthard's work on using corpora to search for clues highlighting anomalies between witness statements and 'verbatim' records has also centred on genre variation (Coulthard 1995), vocabulary choices (Coulthard 1993) and idiosyncratic language (Coulthard 2004). However, as Coulthard acknowledges (1994: 39), the epistemologies and discursive practices of the law courts may be at odds with statistical data from corpora as they operate in the realm of rhetorical persuasion creating opposition to the acceptance of corpus data.

Another problem confronting forensic linguists is the brevity of the texts and transcripts for adducing corpus evidence. Winter's (1997) way of circumventing this drawback relies, not on comparing features of the disputed text with a large-scale corpus such as the Bank of English, but comparing the vocabulary of different texts attributed to the suspect. Winter uses software to select the dominant items of a text, and then scrutinises these to examine their positioning and their co-textual lexico-grammatical environment. In this way, similarities and differences between texts attributed to the same author can be highlighted. Hilton Hubbard (1997) proposes yet another method for authenticating authorship through an error analysis involving identification, description and explanation. While this approach is of value, in principle, for authorship identification its application is in doubt due to the limited size of the corpora available, in the first place, and secondly, the consequent low frequency of errors recorded.

An 'empirically verified theory of idiolect', i.e. a variety of language unique to an individual manifested by patterns of lexis and grammar, has been called for by Kredens (2003) to establish authorship attribution. Kredens (ibid.) carried out an exploratory comparative analysis of speakers with similar biological and social characteristics to investigate patterns of idiolectal variation yielding some promising evidence, which, he states, 'can be used to induce doubt and, as such, could perhaps be utilized to acquit' (p. 442). However, both Solan and Tiersma (2004) and Coulthard (2004: 432) have reservations on whether it is possible to devise a method of *linguistic fingerprinting* given that 'practice is a long way behind theory and no-one has even begun to speculate about how much and what kind of data would be needed to uniquely characterise an *idiolect*'.

6.3.2 Corpora for the analysis of courtroom discourses

Courtroom discourses can be examined through the lens of the judge, the prosecutor, the defendant and their defence, and eyewitnesses.

Quote 6.5 The nature of courtroom discourses

> Courtroom discourses are connected to the 'fact-finding' procedure, which attempts to reconstruct reality through language, e.g. the prosecutor's presentation, the eyewitness's narratives, the defendant's defence and the judge's summing-up. As people may choose to interpret language in different ways according to their own conventions, experiences or purposes, the same word may not mean the same thing to different people. Unsurprisingly, the prosecutor and the defendant produce conflicting accounts of the same event. While the judge's summing-up and the eyewitness's testimonies are supposed to be impartial, studies show that they can also be evaluative.
>
> (McEnery et al. 2006: 117)

This 'evaluative' dimension of courtroom discourses noted by McEnery et al. (ibid.) has been borne out by several studies which make use of corpus data to uncover evaluation through connotation and semantic prosody.

Connotation in the judge's summing-up has been discussed by Stubbs (1996) and in lawyers' cross-examination of witnesses in the courtroom by Cotterill (2004).

Example 6.7 Connotation in the lawyer's cross-examining

Cotterill (ibid.) draws on a 5-million-word subcorpus of UK rape/sexual assault and domestic violence cases (part of a larger 12-million-word corpus of courtroom interaction) to investigate how lawyers exploit connotational meanings, specifically during the dialogic stage of witness cross-examinations, which tend to be adversative in nature. For this, Cotterill consulted the COBUILD Bank of English for verification of connotational nuances.

Cotterill points out that lawyers for the prosecution take on a 'storytelling-by-proxy' role (cf. Cotterill 2003) and are the principal narrators, responsible for topic-raising of the propositions. In the example below, Cotterill shows how the lawyer takes control of the discourse in order to 'defuse the affective force of the connotations' associated with 'lash out', shown to collocate with expressions signifying anger and irrational action in the Bank of English. 'Through this subtle process of neutralization, the lawyer is able to downgrade the severity of the defendant's alleged action, transforming an apparently irrational act of violence – lashing out – into something perhaps more innocuous – a push' (p. 522).

Extract 2 'pushing' vs 'lashing out'

Q. Miss Johnson, would you accept that even on some occasions when the police have been called and the defendants left, literally within hours he is back at your house, your joint house, flat or whatever it is, and you have kissed and made up?
A. This time we never. No. This is when it was all over and the second time the police came was not because of arguing between me and Mr. Jackson. *It's cause he lashed out at my mum.*
Q. *You say 'lashed out'. I think your words are 'he pushed her'.*
A. *Yeah, pushed her.*
Q. *It is slightly different to lashed out, is it not?*
A. *No, because you shouldn't go round pushing people about.*
Q. *Do you not accept there is a difference between a push and a lash?*
A. *Yes.*

(Cotterill 2001: 521)

Semantic prosodies (see Section 1.2.2) have been examined in the prosecution and defence by Cotterill (2003: 68), who found that in the O. J. Simpson trial the repetition of certain words with a negative prosody such as *encounter* deconstructed 'the face of the athlete, the face of the actor', casting Simpson in the role of a 'jealous, possessive husband obsessed by his ex-wife'. Semantic prosody (in this case discourse prosody) has also been ascribed to questioning strategies used in cross-examining. Heffer (2005: 148) notes that one-third of the occurrences of *tell me* are followed by the demonstrative *this*, e.g. *Tell me this: would you accept you ran about sixty yards to get involved in this fight*? Heffer (ibid.) points out that *tell me this* has a negative prosody as it is followed by a proposition which challenges the witness's testimony, implying that the witness may be lying.

Although the corpus approach 'is not likely to achieve the reliability of DNA evidence' (Solan and Tiersma 2004: 461), and more work is needed in the construction of specialist corpora such as text messages, suicide notes, etc. (cf. Cotterill 2010), nevertheless, it has proved its worth in providing linguistic evidence which can be used for consideration in acquittals and which has revealed the type of rhetoric and manipulation of language features prevalent in courtroom discourses. Another branch of the corpus 'family tree' is that of corpus stylistics, an area in which similar phenomena (e.g. semantic prosody) investigated in forensic linguistics are also the subject of enquiry.

6.4 Corpus stylistics research

6.4.1 Role of corpus linguistics in literary stylistics

The branch of linguistics, commonly referred to as corpus stylistics, signifies if not a marriage, at least an engagement, between corpus linguistics and literary stylistics. Before considering how corpus linguistics can be of value in stylistics, it is worthwhile to reflect on how style has been interpreted.

Concept 6.6 O'Halloran on style

> Style is not either inherent in the text (as the formalists claimed) or totally in the reader's mind (as Fish and other reader-response theorists claimed) but an effect produced in, by and through the interaction between text and reader. This meaning and stylistic effect are not fixed and stable, and cannot be dug out of the text as in an archaeological approach, but they have to be seen as a potential which is actualised in a real reader's mind, the product of a dialogic interaction between author, the author's context of production, the text, the reader and the reader's context of perception – where context includes all sorts of sociohistorical, cultural and intertextual factors.
>
> (O'Halloran 2007: 229)

Taking style to represent 'the product of a dialogic interaction' between the author and reader, and by extension literary critic, corpus stylistics can be seen as a systematic attempt to capture snapshots of this 'dialogic interaction'.

Concept 6.7 Corpus stylistics

Mahlberg (2007a: 221) defines the link between corpus linguistics and literary stylistics thus:

> Both are interested in the relationship between meaning and form. Stylistics puts an emphasis on how we say what we say and corpus linguistics also claims that what we say depends on form, i.e. the patterns which are attested in corpora. The focus of the two disciplines, however, tends to be different. Stylistics focuses on what makes a text, or a group of texts, distinctive, and it investigates deviations from linguistic norms that trigger artistic effects and reflect creative ways of using language. Corpus linguistics, on the other hand, mainly focuses on repeated and typical uses that do not only hold in one text, but are found across a number of texts in a corpus. … We can see a link between corpus linguistics and literary stylistics in that 'creativity' can only be recognised as such when there is a language norm against which the 'creative' language comes to stand out.

The two main approaches to corpus stylistics can be said to be either associated with the Lancaster or Birmingham school of 'doing' corpus linguistics (see Section 3.1.1 on probabilistic vs neo-Firthian approaches). Following the Lancaster tradition, corpus linguistics is seen as a methodology which usually relies on annotated text to investigate the kinds of linguistic phenomena described in reference grammars. A case in point here is Semino and Short's (2004) analysis of speech, writing and thought presentation in a corpus of genres which can be broadly considered as 'narrative', consisting of prose fiction, newspaper news reports and (auto) biography. Much of their study is devoted to discussion of their adaptation of the Leech and Short (1981) model for analysing speech and thought in fiction and their revised annotation scheme. Wynne (2006) sees three advantages to this type of approach.

Quote 6.6 Advantages of using annotated corpora in corpus stylistics

> There are typically three outcomes of this process [annotation]. First, the exhaustive analysis of a whole text or corpus is a more empirically sound

(*continued*)

> procedure for discovering linguistic phenomena, compared to choosing examples; annotation of the electronic text forces the analyst to test and refine the system of categorization to account for all cases. Second, it is possible to extract statistics relating to frequency, distribution and co-occurrence of forms from the annotated text. Third, an annotated corpus is obtained, available for studies aiming to replicate or further develop the research, and usable for other areas of literary or linguistic research.
>
> Wynne (2006: 225)

The other approach emanating from the Birmingham school prioritises lexis over grammar and applies the notions of collocation, colligation, semantic prosody and semantic preference in corpus stylistic enquiries. For this reason, it is considered to have its own theoretical context, which goes beyond a purely methodological status. One key advantage of this approach, as argued by Louw (1993, 1997), is that the direct observation method can uncover semantic prosodies associated with particular words which are generally not accessible through human intuition or introspection. In order to detect such kinds of phraseological features, proponents of this more top-down, corpus-driven approach prefer to work with unannotated text so that syntagmatic patterning will not be obscured by, for example, part-of-speech tagging.

The following sections will discuss how various kinds of literary output (novels, plays and poetry) have been analysed using corpus linguistic methodologies, mainly of a comparative nature, to shed light on literary criticism, creative use of language, and stylistic variation.

6.4.2 Literary criticism

Corpus stylistic studies have researched various linguistic features which either confirm and extend the viewpoints of literary critics or, in a few cases, reveal aspects untouched by literary critics or non-aligned with their views.

Culpepper's (2002) analysis of *Romeo and Juliet* reveals striking differences in the keywords, which when examined in context, point to an interpretation of their characterisation, confirming one's intuitions and the viewpoints of literary critics. Culpepper compared each of the six main characters with all the other characters in the play, e.g. the 5000 words spoken by Romeo with the 14,000 words by the other five characters. As Culpepper (ibid.: 15) points out, the choice of reference corpus is crucial in a keywords analysis and 'there is no magic formula for making this decision', a point reiterated by Scott and Tribble (2006).

Example 6.8 Keywords and literary criticism

Culpepper (ibid.: 20) notes the following:

> Romeo's top three keywords 'beauty', 'blessed' and 'love' seem to match one's intuitions about this character: he is the lover of the play. Other keywords, such as 'dear', 'stars' and 'fair' fit his 'love talk' style. For example:
>
> > She hath, and in that sparing makes huge waste; For beauty, starv'd with her severity, Cuts beauty off from all posterity. She is too fair, too wise, too fair. To merit bliss by making me despair; She hath forsworn to love, and in that vow Do I live dead that live to tell it now. (I.i)

Juliet's most 'key' keyword, is the grammatical word 'if', which as Culpepper points out, is not so revealing of characterisation at first glance: e.g.

> If he be married,/My grave is like to be my wedding-bed (I.v.) [at her first sighting of him, whether Romeo is married]
>
> If they do see thee, they will murder thee. (II.ii.) [whether Romeo will be spotted during a covert visit]

However, when considered along with other keywords such as 'yet', a picture emerges indicating Juliet's state of anxiety throughout the play:

> Tis almost morning; I would have thee gone; And yet no further than a wanton's bird [..] (II.ii.) [whether Romeo should go]

Culpepper maintains that 'if' and 'yet' create a style that is meaningful; it articulates Juliet's anxieties. This style is supported by other keywords, such as the subjunctive 'be' …and the modal 'would' …, where it expresses her mixed wishes' (pp. 20–1).

Likewise, Mahlberg (2007b, 2010) uses cluster analysis to identify phrases which point to the characterisation noted by literary critics.

Example 6.9 Functional clusters and literary criticism

Mahlberg (2007b: 20) notes that literary critics have pointed out the particular function of the 'fanciful as if' in Dickens. She cites Brook (1970: 33) who notes that 'It generally takes the form of the invention of some improbable but amusing explanation of the appearance or behaviour of one of the characters in a novel.' The importance of a suitable reference corpus is to be

(*continued*)

emphasised here. By comparing the 4.5-million-word Dickens Corpus with a comparable sized corpus of nineteenth-century fiction (see Table 6.1 below), Mahlberg is able to isolate key 'as if' clusters which are specific to Dickens's works. An examination of the wider context of these 'as if' clusters was then carried out to determine if they fulfilled this function.

Table 6.1 'As if' clusters (from Mahlberg 2007b: 15)

Number of types: 5	***Dickens Corpus***	***19th Cent.***
as if he would have	41	2
as if he were a	45	7
as if he were going	32	3
as if it were a	72	23
if he were going to	26	3

Meanwhile, Starcke's (2006) and Fischer-Starcke's (2009, 2010) research on the literary works of Jane Austen also provides evidence on how corpus linguistic techniques can both enhance and *extend* appreciation of characterisation and plot.

Example 6.10 Phraseologies and literary criticism

In her corpus stylistic analysis of Jane Austen's *Persuasion* (around 83,400 words), Starcke (2006) first extracted the top sixteen 3-grams, with *she could not* found to be the most frequent, occurring 54 times. Her more in-depth analyses both serve to confirm the previous intuitions of literary critics and to detect aspects of the novel hitherto unnoticed. For example, Starcke (ibid.) notes that literary critics perceive one of the main themes of the novel as the passing of time and the consequent fading of beauty of its main protagonist, Anne Elliot, who broke off her engagement to Captain Wentworth on her family's advice that she should seek a more wealthy husband. This intuitive perception is borne out by the concordance output for *she could not*, which creates a link between Anne and the main theme (see Table 6.2).

Table 6.2 *She could not* with indicators of time (Starcke 2006: 93)

enant." Ann could listen no longer;	she could not	Even have told how the po
nting him. This was almost cruel. But	she could not	be long ungrateful; he wa
ave said, that "now Miss Anne was come,	she could not	suppose herself at all wa
H they had now been acquainted a month,	she could not	be satisfied that she rea
the prospect of an hour of agitation	she could not	quit that room in peace wi

Starcke (ibid.) remarks that the melancholic atmosphere of the novel has been ascribed by critics to the time of year when the novel is set and the unhappy life Anne leads with her family. However, Starcke's analysis reveals that references to time frequently co-occur with grammatical negatives and words with a negative connotation, which are linked to an agent, Anne. Starcke thus makes a linguistic link between Anne and the sombre, melancholic atmosphere of the novel, noting that 'The fact that it [the atmosphere] is part of the language and closely linked to a single character has been frequently overlooked by critics of the novel' (p. 98).

Stubbs's (2005) analysis of Joseph Conrad's *Heart of Darkness* pinpoints grammatical features that critics have disregarded in their discussion of one of the major themes of this book regarding the colonist Marlow's 'unreliable and distorted knowledge' in his narration of his travels up a river in Africa in search of an ivory trader (p. 9).

Example 6.11 Grammar and literary criticism

Stubbs (2005: 9) acknowledges that critics are in agreement on the vague impressions and unreliable knowledge in Marlow's narration, summarising how their interpretation rests on the image of mist and haze and words from the lexical field of darkness, with a total of 150 items averaging one per page:

- blurred 2, dark/ly/ness 52, dusk 7, fog 9, gloom/y 14, haze 2, mist/misty 7, murky 2, shadow/s/y 21, shade 8, shape/s/ed 13, smoke 10, vapour 1

Stubbs's analysis reveals, though, that the above list of linguistic features is somewhat incomplete:

> However, they [critics] tend to ignore the many grammatical words denoting vagueness and uncertainty. The word *something* occurs over 50 times, in expressions such as:
>
> - I don't know – <u>something</u> not quite right
> - Reminded me of *something* I had seen – something funny
>
> There are over 200 occurrences of *something, somebody, sometimes, somewhere*, *somehow* and *some*, plus around 100 occurrences of *like*

(*continued*)

(as preposition), plus over 25 occurrences of *kind of* and *sort of*, all often collocated with other expressions of vagueness:

- the outlines of some sort of building
- seemed somehow to throw a kind of light
- I thought I could see a kind of motion
- indistinct, like a vapour exhaled by the earth…misty and silent

(Stubbs 2005: 10)

One study that casts some doubt on critics' established views of characters is that by Carretero González and Hidalgo Tenorio (2005). As indicated below, their exploratory corpus study of modality markers in *Hamlet* does not reflect 'the most accepted view of the Prince of Denmark as an irresolute, passive procrastinator' (p. 1).

Quote 6.7 Modal verbs and literary criticism

The cases in which Hamlet's willingness to be a man of action is explicitly manifested outnumber those in which his lack of certainty or his sense of duty is marked by modal verbs. Both in his interaction with other people in the Court and in his soliloquies, Shakespeare's creature is depicted as someone who is eager to perform whatever he is supposed to do. In four of the five acts of this play, those modals indicating volition and inclination are the ones Hamlet comes to prefer (volition including wanting, wishing and intending, sentences that express promises and threats are also volitional). This might mean that we should focus our attention on his *desire* to be an agent rather than on his *doubts* to be so. Furthermore, in his characteristic use of modals, Hamlet also prefers those conveying logical necessity, which do not primarily involve human control over events but do typically involve human judgment of what is or is not likely to happen.

(Carretero González and Hidalgo Tenorio 2005: 5)

6.4.3 Creative use of language

While the studies cited above usually make use of literary texts by the same author or from the same time period to shed light on characterisation and plot, creative use of language is often investigated through comparison with a reference corpus of non-literary texts.

Louw's (1993) seminal work on collocations and semantic prosodies in literary texts has been instrumental in laying the foundation for investigation of literary devices. Hoey's (2005, 2007) theory of lexical priming (see Concept 1.11) also has relevance for creativity in literary works, as illustrated below.

Concept 6.8 Hoey's theory of lexical priming in relation to creative language

According to Hoey, literary creativity is achieved when primings are overridden, as in cases of collocational clashes (cf. Louw 1993). As illustration, Hoey (ibid.: 177) takes the first two lines of Dylan Thomas's poem *A grief ago*:

> A grief ago
> She who was who I hold, the fats and flower,

Using statistical evidence from a corpus of newspaper texts, Hoey's analysis of *ago* shows it to be primed for collocation with *years, weeks* and *days*, for semantic preference with units of time, e.g. *six weeks ago*, for colligation with adjunct function, and for text-initial position, when it is sentence-initial, among others. Hoey explains how Dylan Thomas's use of *ago* breaks some of the primings, i.e. collocational and semantic, whilst adhering to the other sets of primings, for literary effect (Hoey 2005).

Louw's example below shows how the semantic prosody of an item, i.e. 'a consistent aura of meaning with which a form is imbued by its collocates' (p. 157), is contravened in a literary text to achieve an ironic effect.

Example 6.12 Detecting irony through examination of semantic prosody

Louw (1993: 164–5) asks the reader to consider the following short passage from the novel *Small World* by David Lodge:

> The modern conference resembles the pilgrimage of medieval Christendom in that it allows the participants to indulge themselves in all the pleasures and diversions of travel while appearing to be austerely *bent on* self improvement. (emphasis added)

Louw's search on *bent on* in the original 18-million-word COBUILD corpus shows there to be 10 citations for *bent on*, 7 of which have a negative semantic prosody.

them were so	bent on	defending themselves and on distinguishin
ment is hell	bent on	destroying British Leyland, aided and abe
side, seemed	bent on	getting down my collar and up the trouser
Of the crowd	bent on	harrying the speakers, often for a laugh
nt. They are	bent on	'improving' and perfecting existing weapo
N or persons	bent on	mayhem had not so far chosen to resort to
ated figure,	bent on	on the same routine. Thereafter every Mamous

(*continued*)

The remaining three instances, e.g. 'resolutely bent on the rescue to which they had been called', were all drawn from the same literary source.

Thus, through investigating the typical semantic prosody of *bent on* in the general COBUILD corpus, Louw is able to show how David Lodge creates the desired ironic effect.

To return to the novels of Dickens, Hori (2002) has investigated unfamiliar collocations, i.e. deviant from the norm, of types referred to as figurative and hybrid, as summarised below.

Example 6.13 Deviant collocations in Dickens

Figurative collocation
In this type of collocation, a manner adverb modifies an adjective or verb grammatically but functions in a figurative way: e.g.

> 'The Commandments say, no murder. NO murder, sir!' proceeded Mr. Honeythunder, *platformally pausing* as if he took Mr. Crisparkle to task for having distinctly asserted that they said: You may do a little murder, and then leave off. (*The Mystery of Edwin Drood*, Ch. 17)

Hori draws attention to the phrase *platformally pausing* conveying the affected pose of the hypocritical philanthropist as if standing on a speaker's platform.

Hybrid collocation
This type of collocation is oxymoronic in nature. In the example below, there is a contradiction between appearance and reality.

> 'Eh?' The Father of the Marshalsea always lifted up his eyebrows at this point, and became *amiably distraught* and *smilingly absent in mind*. (*Little Dorrit*, Ch 18)

Hori points out how combinations of semantically oxymoronic words referring to William Dorrit's expression and mind, contribute to producing humour, irony and characterisation (Hori 2002: 154–9).

Hori (ibid.: 157) also posits another type of collocation, subject-oriented, which refers to a manner adverb that semantically qualifies the subject in addition to the verb, as in the following from *Great Expectations*: '"No, Wegg," said Mr. Boffin, *shaking his head good-humouredly*'. Although the reader might

consider this collocation as somewhat unusual, it does not seem deviant in the same way that figurative and hybrid collocations are. Carter (2004) does not see a stark difference between literary and non-literary language, but rather views literary language as a continuum. The three types of collocations investigated by Hori would seem to represent different degrees of literariness along the cline proposed by Carter.

Symbolism and leitmotifs are also realised through collocations and semantic prosodies, and, importantly, conveyed via *recurrent* collocations. Their instantiation in text can be viewed as overriding the normal primings, as discussed above with reference to Hoey's work. Louw (1997) illustrates how a symbol is developed through an accumulation of unusual collocations in the work of Sylvia Plath.

Example 6.14 Symbolism developed through recurrent collocations

… in the case of Sylvia Plath the words *eyes* and *bald* collocate in ways that contribute to her symbolism. *Eyes* are normally associated with ways of reading the emotions of other people. However, for Plath they are rendered expressionless by the form *bald*, a collocate which nowhere appears near the form *eyes* in the entire Bank of English. …This symbol, as Plath has decided to develop it, has an obvious influence in establishing the strong sense in her work that relationships which purport to be close are, in reality, dysfunctional (see Figure 6.2).

1. The *bald*, white tumuli of your eyes
2. She may be *bald*, she may have no eyes, She's pink, she's a born midw
3. Nor leave me to set my small *bald* eye Skyward again, without hope of,
4. ed rocks sunning in rows, *Bald* eyes or petrified eggs, Grownups
5. estone The *bald* slots of his eyes stiffened wide open On the inc
6. ight around my bed, Mouthless, eyeless, with stitched bald hea

Figure 6.2 Sample concordance on 'bald' in Sylvia Plath (Louw 1997: 248)

The works of German novelists, specifically Goethe and Thomas Mann, have also been the subject of corpus-based enquiries looking at symbolism and leitmotifs (cf. Burgess 2000; Lawson 2000). For example, Burgess (ibid.) used corpus linguistic techniques to pick out the recurring images of *the* glass and glass, which links the two protagonists in the novel, Eduard and Ottilie, from its beginning to end.

Example 6.15 Leitmotif developed through recurrent collocational clusters

Burgess (ibid.) refers to the leitmotifs discussed by literary critics in Goethe's *Die Wahlverwandtschaften*, one of which relates to glass. A search for the terms 'Kelch' (cup) and words beginning with 'Glas-' or 'Gläs-' yields results which can all be traced back to Eduard in some way.

A search for 'Kelch', 'Glas-' and 'Gläs-' (adapted from Burgess 2000: 49)

101	er ein wohlgeschliffenes	**Kelchglas**	auf einen Zugaus und warf es
101	…hnitten: es war eins der	**Gläser**,	die für Eduarden in seiner Jugend
255	…rgen, indem sie in ein	**Glas**	wein blickt, das sie eben auszuschlürfen
256	aus dem durchsichtigen	**Glase**,	Worin sich, ob sie gleich zu trinken
414	…ung scheint er aus dem	**Glase**	Zu schlürfen, das ihm freilich kein

Burgess explains that the two examples on pp. 255 and 256 refer to the glass from which the 'mother' is drinking in a *tableau vivant*. He notes that the verb 'schlürfen' (slurp) is used only twice in the novel, in the form of 'auszuschlürfen' on p. 255 for the 'mother', remarking 'somewhat strangely, surely, for a supposedly respectable middle-aged lady! … In contrast to Eduard's own immoderate drinking habits, here the wine does not diminish in her glass, however long the 'mother' seems to be drinking' (p. 50). The leitmotif of the glass is taken up throughout the novel. It is introduced at key events such as the throwing of the glass at the combined celebration of Ottilie's birthday party and the foundation stone laying, and again at the end of the novel when Eduard insists that Odile is laid to rest in an open coffin which is sealed with a glass lid (Burgess 2000).

6.4.4 Stylistic variation

Another author whose works have attracted the attention of linguists is Henry James. Using an annotated corpus, Moss (2009) has compared patterns of syntactic complexity between Henry James's early novel, *Washington Square*, and a late novel, *The Golden Bowl*. Comparisons are also made within each novel and references made to literary criticism.

Example 6.16 Using an annotated corpus to investigate syntactic complexity

Moss (2009) notes that various critics have commented on the 'difficult writing' of Henry James, which becomes increasingly complicated in his later novels. However, Moss states that this complexity in writing has only tenuously been related to syntax by literary critics.

Moss's starting point for this project is with some modification to existing annotation software, the International Corpus of English Corpus Utility Program (ICECUP) developed under the auspices of the Survey of English Usage at University College London. While the existing software was used to tag her corpus with part of speech and function labels and also tag features such as transitivity, passivisation and inversion of verbs, she also added elements of parsing such as categorisation of various clauses.

The initial findings indicate that within a single novel syntactic complexity appears in waves throughout the text, with less complex syntax found in sentences with direct speech. Preliminary findings also show an increased complexity in syntactic structure between the early and the later novels, but that critics' impressions may be based on a few idiosyncratic sentences.

A somewhat unusual application of computerised stylistic variation is that for diagnosis of a disease. Changes were tracked in the writing style of Iris Murdoch, who suffered from Alzheimer's in the last few years of her life, to determine the effects of the disease on the brain's semantic system. This was a collaborative project between UCL and the Medical Research Council's cognition and brain sciences unit in the UK.

Concept 6.9 Using stylistic variation to track the effects of Alzheimer's

Three of Iris Murdoch's novels written at different stages of her life were compared: *Under the Net* (1954); *The Sea, The Sea* (1978), written during the prime of her creative powers, and *Jackson's Dilemma* (1995), published the year before her diagnosis.

For each novel, statistics were compiled of the number of different word types, together with their number of tokens. To show how her use of language compared with average use, word types were weighted with an estimate of how frequently each one appeared in general written usage. The smallest number of word types occurred in *Jackson's Dilemma* and the largest in *The Sea, The Sea*. *Jackson's Dilemma* was found to contain the most commonplace vocabulary and *The Sea, The Sea* the most unusual. The richest vocabulary was also found in this novel which had a smaller number of word tokens per type.

Results from this type of interdisciplinary research project could prove to be useful for the design of cognitive tests to diagnose the disease at the earliest possible stage.

(Dr Peter Garrard, cited in Highfield, the *Daily Telegraph* 2004)

6.4.5 Cautions and future directions

This section has surveyed the diverse range of projects being carried out in corpus stylistics, the majority of which follow the Sinclairian corpus-driven approach. Annotation of literary corpora has also proved of value in this field. These studies have shown how literary appreciation can be enhanced through a corpus-analytic perspective in that corpus evidence can be used to corroborate, extend, or even in a few cases, call into question some traditionally accepted evaluations. In fact, both Stubbs (2005) and O'Halloran (2007) have argued that the systematicity of corpus-based enquiries and corpus evidence can to some extent offset the criticisms by Stanley Fish that 'stylistic analysis is arbitrary and circular' (O'Halloran 2007: 227).

Although there has been some engagement between corpus linguistics and literary stylistics, the relationship is for the moment somewhat one-sided. For its part, corpus stylistics has been criticised for being reductionist in nature and promoting superficial reading of text with its focus on surface forms (see Quote 7.14). For the time being, corpus linguists seem more enthusiastic about what their methodologies and corpus-analytic theories of phraseology can bring to the field of literary appreciation; so far, literary critics do not seem to have embraced this corpus-orientation to the study of literary texts. However, corpus linguists are mindful of this dissonance between the two fields, with Louw (2006) and Archer (2007) sounding a note of caution and stressing that interpretation of literary texts relies essentially on the analyst, and tools are but an aid in this process.

6.5 Translation studies research

Baker (1993, 1995, 1996), a pioneer in introducing corpus linguistic methodologies to translation studies, notes the paradigm shift in the theoretical framework of machine-based translation studies triggered by the use of corpora. This watershed signifies a move away from formal and conceptual representations of language with a reliance on introspective methods emphasising the notion of a one-to-one semantic equivalence between the source text and the target text towards a more socio-cultural perspective.

Quote 6.8 Baker on developments in translation studies

> The move away from source texts and equivalence is instrumental in preparing the ground for corpus work because it enables the discipline to shed its longstanding obsession with the idea of studying individual instances in isolation (one translation compared to one source text at a time) and creates a requirement which can find fulfilment in corpus work, namely the study of large numbers of texts.
>
> (Baker 1993: 237)

Hatim (2001) also remarks on this more descriptive framework informed by sociocultural considerations for translation studies, noting its alignment with the field of corpus linguistics.

Quote 6.9 Corpus linguistics and corpus translation studies

A number of concerns are shared by the two fields of enquiry (corpus linguistics and corpus translation studies):

- Primacy is accorded to authentic instances of language use, and to a move away from introspection (Holmes 1978: 101).
- Texts are viewed not as idealized entities but rather as observable facts (Toury 1980: 79).
- A concern with what corpora should consist of and with how to guard against such pitfalls as bias in the selection of materials.
- Recognition that computational and statistical tools are not sufficient by themselves, and that intuition and observation have a role to play.

(Hatim 2001: 82)

6.5.1 Types of translation corpora

Most translators working with corpora have identified three main types of corpora: parallel, comparable and multilingual. However, in reality there is some overlap among all three categories and inconsistency in the literature in defining these. Of course, monolingual corpora also have a role to play in translation studies, as pointed out by Bernardini et al. (2003: 6). They can 'help (future) translators opt for natural 'native-like' turns of phrase, appropriate to the communicative situation in which the target text will be operating'.

Concept 6.10 Definition of parallel corpora

Parallel corpora contain texts and their translations into one or more languages. A bilingual parallel corpus contains texts and their translations into one language, and a multilingual parallel corpus contains texts and their translations into two or more languages. Thus, a parallel corpus of computer user documentation, for example, will contain the original documents and their translations into one or more language(s). However, while the word 'parallel' is used to indicate that a corpus contains texts and their translations, the text pairs in a parallel corpus are not always translations of each other. They can be translations of a third text. For example, you might have a bilingual parallel corpus containing the French and German translations

(*continued*)

> of computer user documentation produced in English by an American company. In such circumstances, the user is likely to know that both sets of texts are translations. There are, however, situations where the user has no idea in which language a particular set of texts was originally written. This often happens in multilingual environments (in the European Union, for example) where there is more than one official language.
>
> (Bowker and Pearson 2002: 92–3)

Probably the best known parallel corpus is the Canadian Hansard corpus containing records of Canadian parliamentary proceedings in both French and English. It is classified as bidirectional as the source texts can be in French and their translations in English, or the source texts can be in French with translations in English. Another parallel corpus is COMPARA, a bidirectional corpus of English and Portuguese (cf. Santos and Frankenberg-Garcia 2007).

Comparable corpora differ from parallel corpora in that the texts are not translations of each other, but can be compared for similar characteristics across several dimensions. However, to what extent a comparable corpus can be said to be truly comparable, especially in terms of balance (Baker 2004), is open to debate (Laviosa 1997).

Concept 6.11 Definition of comparable corpora

> The shared features [of comparable corpora] will frequently include subject matter or topic and may also include features such as text type, period in which the texts were written, degree of technicality etc. An example of a comparable corpus would be a set of research papers (shared text type) in two or more languages dealing with genetic engineering (shared subject field), written in the last twenty years (shared period).
>
> (Bowker and Pearson 2002: 93)

6.5.2 Corpora and translation universals

Corpora have proved of value in investigating translation from the perspective of 'translation universals', which are features inherent in translation per se and independent of the source language variables.

Concept 6.12 Universals of translation

> *Simplification*
> The notion that translators subconsciously simplify the language or message or both

Explicitation
The tendency to spell things out in translation, including, in its simplest form, the practice of adding background information

Normalisation
The tendency to conform to patterns and practices that are typical of the target language, even to the point of exaggerating them

(Adapted from Baker 1996: 176–7)

These three universals of translation have been the subject of various corpus studies, as outlined below.

Simplification

Simplification can involve strategies such as the breaking up of complex sentences, omissions of repeated information, shortening of collocations, the use of superordinate terms instead of hyponyms when no equivalence is available in the target language, etc. Simplified vocabulary range and choice has been the subject of several corpus-based investigations. Laviosa (2002), using the TEC (Translational English Corpus) and NON-TEC comparable corpus, tested lexical simplification by examining the range of vocabulary used and the lexical density of the two subcorpora.

Example 6.17 Laviosa's findings on simplification

Range of vocabulary
The translational corpus was found to have a narrow range of vocabulary, confirmed in three ways. The proportion of high frequency words to low frequency words was higher in the translational corpus. The 108 most frequent words (frequency list head) in the translated texts accounted for a larger proportion of the corpus than the frequency list head of the corpus on non-translations. The list head of the translation corpus contained fewer lemmas.

Lexical density
Lexical density, calculated by measuring the proportion of content words to grammatical words, was found to be lower for the corpus of translations.

(Laviosa 2002)

However, Xiao (2010) queries whether translation universals, so far investigated with respect to European languages, can be generalised to non-European

languages such as Chinese. Taking Laviosa's work as a baseline, Xiao investigated features of simplification such as lexical density and sentence length using the ZJU Corpus of Translational Chinese, created specifically to study the features of translated text in relation to non-translated Chinese in the Lancaster corpus of Mandarin Chinese representing native Mandarin Chinese. Laviosa's findings on lexical density were supported by Xiao's analysis (i.e. translational Chinese was found to have a significantly lower lexical density than native Chinese), while other aspects such as sentence length were deemed to require further analysis as they were found to be genre sensitive. Xiao's research has thus opened up exciting possibilities for other lines of enquiry.

The issue of vocabulary simplification has also been investigated from a longitudinal perspective. Utka (2004) took a process-oriented approach to this phenomenon by looking at three phases of translating, the first translator's draft, the second edited draft and the final version of translation, in an English–Lithuanian corpus of European community law documents. The analysis of missing types through 'tracing a complete removal of a word form or a replacement of one word form by another in the later editing stages' revealed instances of simplification (p. 208). For example, it was found that five English words (*wastes, sludges, slag, muds, dross*) were gradually reduced just to one Lithuanian word (*dumblai*) in the final version. Moreover, Utka emphasises that the search for missing types in his three-phased translation corpus is a useful method for finding cases of translationese (interference from the source language), e.g. unnatural compounds translated on a word-by-word basis, as such features 'tended to be replaced systematically in later editing phases' (p. 210).

Explicitation

Studies on explicitation have researched optional and obligatory aspects of this phenomenon. Olohan and Baker (2000) and Olohan (2003) investigated the omission of optional syntactic features in TEC and the BNC, finding support for the notion of explicitation in the corpus of translations. Also see Frankenberg-Garcia's (2009) study on explicitation investigated from the perspective of text length, but also taking into account obligatory explicitation dictated by the grammar of the source language, i.e. Portuguese.

Example 6.18 Inclusion of optional syntactic features as evidence of explicitation

*Omission of modal **should** from a THAT complement*
As Table 6.3 shows, a greater proportion of omission is seen in BNC.

Table 6.3 ORDER and SUGGEST + that + should/zero in BNC and TEC

Form	BNC	TEC
ORDER + *that* + *should*	1	6
ORDER + *that* + zero	2	7
SUGGEST + that + *should*	19	19
SUGGEST + that + zero	43	58

However, Olohan advises caution as inclusion or omission of syntactic features may be affected by other co-occurrence features such as use of modifiers, pronominal forms, modal constructions, etc.

(Olohan 2003: 429)

Normalisation

Normalisation, sometimes referred to as conventionalisation, has been the focus of several studies in the domain of literature. Bosseux (2004) brings up the question of 'conventionalisation' in her study of three different French translations of Virginia Woolf's novel *To the Lighthouse*. Focusing on linguistic features such as modality, ergative and transitive constructions constituting the notion of 'point of view', she calls into question whether a translated text can be truly conventionalised in terms of style, pointing out that the translator's discursive presence will always be felt in the text.

6.5.3 Corpora and creative use of language

The translation of creative use of language has also been the subject of discussion and research. Whether more conventionally encoded language is found in translated target texts than the source texts has been investigated in various types of literary works with respect to the creative use of lexis (Malmkjær 1998, 2009), and collocations, semantic prosody and preference (Kenny 1998, 2000). Malmkjær (1998) notes that in a parallel corpus of multiple translations into English of one Danish source text, one of Hans Christian Andersen's fairy tales, the majority of translators produced the lexically conventional translation *The Princess and the Pea*, in spite of the fact that creative language use was found in the source text, i.e. *The Princess on the Pea*. Both Malmkjær (2009) and Kenny (1998) seem to agree that creative use of language should be mirrored by the translation of the target text: 'an unconventional ST wording should be matched by an equally unconventional TT wording' (Kenny 1998: 5). In this kind of situation, Malmkjær (1998) argues that using a parallel corpus of previous translations of this story or even large standard corpora may be misleading as past language behaviour cannot be assumed to be a model for future linguistic behaviour, and that creative inspiration and introspection also have

a role to play. Kenny, meanwhile, advocates using reference corpora on the grounds that if a certain collocation is not found then it could be deemed to be unusual.

Example 6.19 Using reference corpora to verify unusual collocations

Kenny (1998: 3) discusses the collocation 'three-cornered glance' from the novel *Digging to Australia* by Lesley Glaister. She notes that the translator preserved an unconventional compound, *Dreiecksblick* (literally a 'triangular glance') in the translation into German:

> Bob's eye's flickered over me, a brief three-cornered glance, a check for progress. I felt like some sort of time-bomb. (Glaister 1992: 38)

> Bob musterte aus den Augenwinkeln, ein kurzer Dreiecksblick, ein Abschätzen meiner Fortschritte. Als ob ich eine Zeitbombe wäre. (Glaister 1995: 54)

Kenny suggests that it would be useful for translators to consult reference corpora for cases like this. A search in COBUILD showed there to be an association between the word 'glance' and the modifiers 'sidelong', 'sidewise', 'cursory', 'furtive', 'upward', 'backwards', but 'three-cornered' was only used with 'contests', 'fights' and the 'hat' of Massine's ballet. Thus consultation of a reference corpus in this way can ensure against normalising creative collocations.

(Adapted from Kenny 1998)

Translation studies research has also been discussed with reference to 'translationese', which can be due to the influence of the source language on the target language, or due to 'strategies of avoidance or overindulgence' (Schmied and Schäffler 1996: 48). The latter strategies might involve greater explicitness or simplification, thus overlapping to some extent with translation universals. The former strategy relates to the concept of contrastive interlanguage discussed in the following chapter on learner corpora.

6.6 Learner corpora and SLA research

In a 2002 survey article Pravec noted that learner corpora are mostly found in Europe and Asia, and this is a situation which still predominates today. The reason for this is largely historical as these regions happened to be sites

where learner corpora were first launched. An extensive learner corpus project, the International Corpus of Learner English (ICLE), was initiated by Sylviane Granger at the University of Louvain, Belgium, in 1990. This project parallels the large-scale ICE (International Corpus of English) project encompassing different varieties of English (see Section 5.3.2). ICLE consists of subcorpora of academic argumentative essays written in English by French, German, Polish, Greek, etc. advanced learners. In the early 1990s, a learner corpus comprising academic writing was established by John Milton at HKUST (Hong Kong University of Science and Technology), while in Japan in the mid-1990s Yukio Tono launched the JEFLL (Japanese English as a Foreign Language Learner) Corpus consisting of academic texts produced by students at the junior high, high school and university level.

While written corpora predominated in the early stages of learner corpora compilation, spoken corpora have now also come on stream (see case study in Section 8.1). Other developments in learner corpus research include a branching out into contrastive interlanguage research and the exploration of learner corpora for the study of language acquisition. However, in the main, learner corpora have mostly been used to compare native vs non-native speaker usage, or from a slightly different perspective, apprentice vs expert language, as discussed below.

6.6.1 Learner corpora: interlanguage features

Learner corpora are usually tagged for errors (see Dagneaux et al. 1998), which can aid subsequent analysis. This usually involves looking at the factors which may be the origin of interlanguage features, as outlined below.

Quote 6.10 Explanations for interlanguage features

L1 transfer
Some forms or grammatical patterns found in the learner's language production may result from the intrusion of L1.

General learner strategies
To help deal with the complex task of speaking or writing in a second language, the learner may adopt some coping strategies such as the use of L1 forms, circumlocution, avoidance strategies, etc.

Paths of interlanguage development
Some aspects of interlanguage, such as the development of negation or the development of tense/aspect marking proceed in a series of stages which may be tracked using longitudinal studies of learner output.

(*continued*)

Intralingual overgeneralization
Some features of the learner's language may be due to overgeneralization of an aspect of L2 grammar such as the use of *-ed* to mark past tense.

Input bias
The form of the learner's production may reflect the particular input received, such as the language used in coursebooks (see Römer 2004).

Genre/register influences
Researchers working with learner corpora have suggested that the writing of L2 learners contains a variety of informal patterns that are characteristic of spoken discourse.

(Barlow 2005: 343)

The above explanations for various interlanguage features will be taken up in the following sections discussing major written and spoken learner corpora.

6.6.2 Learner corpora: written

Different types of interlanguage features have been attributed to L1 transfer (also see Section 6.6.4). For example, Granger (1998b) noted the overuse of the formulaic sentence builder *we/one/you can/cannot/may/could/might say that …* in the ICLE component of argumentative writing by French learners of English, which she ascribes to L1 transfer as French uses many more phatic introductory phrases than English. Nesselhauf (2003, 2005) also remarks on possible L1 transfer in her study of advanced learners' use of collocations in the German sub-component of ICLE, assuming that the collocation **make homework* was influenced by the L1 as 'German has *Hausaufgaben machen* and that *machen* is related to *make* in both meaning and form' (Nesselhauf 2003: 234).

Two studies linking epistemic modality to input bias are those by Hyland and Milton (1997) and McEnery and Kifle (2002). In the study by Hyland and Milton the lack of hedging devices in the academic essays of Hong Kong undergraduate students was ascribed to the emphasis on the teaching of such expressions as 'There is no doubt that …' at tutorial schools in Hong Kong. On the other hand, McEnery and Kifle found an overuse of hedging devices which could be accounted for by the learning of lists in secondary school coursebooks. Another study related to hedging is that by Lorenz (1998: 62), who found overuse of adjective intensification in the writing of German learners, e.g. *A really visible reason for the emancipated woman being alive is the high rate of unmarried or divorced women*. His observation that these writers are putting too much weight into the theme of the clause, violating 'the principle of end-weight', seems to point to a lack of teaching on information structure. Meanwhile, Leńko-Szymańska (2004) attributes advanced Polish students' lack of mastery of demonstratives

as anaphora markers to the fact that teachers and learners tend to view them as trivial features, calling for explicit instruction in this area (see McCrostie 2008 for a positive effect of teaching input on learner writing).

While a substantial amount of research has been carried out on argumentative essays of learner academic writing, a wider range of written academic genres is now under compilation with the British Academic Written English, BAWE, corpus and a similar corpus, the Michigan Corpus of Upper-level Student Papers, MICUSP. One issue that has surfaced concerns classification due to the increasing interdisciplinarity of academic subjects, e.g. psychology and the law, physics in medicine (Nesi 2011).

These two initiatives are impressive undertakings, but whether corpora such as BAWE and MICUSP should be considered as 'learner' corpora is somewhat of a moot point, though. On the one hand, they can be considered as 'expert' on account of the fact that all assignments are top-rated ones in terms of content. And, indeed, Wulff and Römer's (2009) contrastive research on MICUSP indicates that the writers show mastery of academic genres at the phraseological level (see Callies 2009 for further discussion of this issue). On the other hand, BAWE and MICUSP could also conceivably be seen as 'learner', regardless of whether the students are L1 or L2 writers, as obtaining high scores may not necessarily equate with genre mastery and in some cases there may be phraseological infelicities. Tentative evidence for this point from a pilot study on authorial stance using the BAWE corpus (Henderson and Barr 2010: 261) indicates that native and non-native students alike 'are not yet ready to take on the status of the field expert who evaluates others' work'. More corpus research and consultation with content lecturers would be necessary to verify whether absence of this type of expertise is acceptable, or whether it is to be regarded as somewhat of a deficiency.

6.6.3 Learner corpora: spoken

The majority of research undertaken on spoken corpora has involved the Louvain International Database of Spoken English Interlanguage (LINDSEI) and the native control corpus, i.e. Louvain Corpus of Native English Conversation (LOCNEC). For example, De Cock's (2000) investigation of the use of formulae by French speakers, showed that some such as *for example, of course* were overused, while others, e.g. *sort of, you know*, were underused. Formulaic language is one indicator of fluency, as implied by De Cock's study. Other corpus studies on the French component of LINDSEI have investigated fluency in terms of speech management strategies involving phenomena such as repeats, self-repairs, discourse markers, hesitations, etc. (Gilquin 2008), or temporal fluency variables involving speech rate, length of speech or the length and number of pauses (Osborne 2011). The pilot study by Brand and Götz (2011) on the German component of LINDSEI is noteworthy for its application of a

multi-method approach to examine a possible correlation between accuracy, defined in terms of lexical, grammatical and phonological errors, and temporal variables of fluency.

Discourse markers (DMs), a type of speech management strategy, have been investigated in a pedagogic subcorpus from CANCODE (see Section 3.4) and a corpus of interactive classroom discourse of secondary school students in Hong Kong (Fung and Carter 2007).

Concept 6.13 Fung and Carter's taxonomy for investigating DMs

In Fung and Carter's taxonomy DM are divided into four categories: referential (indicating relationships between utterances); structural (showing how sequences are organised and managed); interpersonal (marking shared knowledge) and cognitive (denoting thinking processes), as outlined below.

<2>	**Right**. (laughs)	[structural/interpersonal]
<2>	**And** and again that was it was sharing the same subject.	[referential]
<1>	**So** you've got coordinated clauses there.	[structural]
<2>	**Yeah**. Mm. I've got one there as well.	[interpersonal]
<1>	**Right**.	[interpersonal]
<2>	**I think** it's	[cognitive]
<1>	**Right**.	[interpersonal]
<2>	Em there's a mad sentence here. which Pause	

(Adapted from Fung and Carter 2007: 428)

Of interest is that while native speakers were found to use discourse markers for a wide variety of pragmatic functions, Hong Kong students used referentially functional discourse markers (*and, but, because*, *OK*), but displayed a relatively restricted use of interpersonal markers (*sort of, you know*). That these data have some overlap with De Cock's (2000) findings suggests that these are generalisable interlanguage features, another aspect of learner corpus research discussed in the following section.

6.6.4 Contrastive interlanguage analysis

As Gilquin (2000/2001) points out, contrastive interlanguage analysis (CIA) can be combined with pure contrastive analysis (CA) in a model that has been called the integrated contrastive model, an area first highlighted by Granger (1996).

Concept 6.14 Contrastive interlanguage analysis

> Contrastive interlanguage analysis (CIA) involves quantitative and qualitative comparisons between native language and learner language (L1 vs L2) and between different varieties of interlanguage (L2 vs L2). The first type of comparison plays an important role in uncovering the distinctive features of learner language, while the second makes it possible to assess the degree of generalisability of interlanguage features across learner populations and language situations.
>
> (Granger 2009: 18)

Gilquin and Paquot's (2008) study on the overuse of *let's/let us* in French learners' essays can be traced back to the L1 (see Example 6.20). However, at the same time they caution against ascribing L1 transfer as the sole explanation for the overuse of this feature. They note that other factors must come into play as this imperative form is also used by learners from other mother-tongue backgrounds (e.g. Dutch) which do not make use of first person plural imperative verbs. They thus suggest other reasons for this phenomenon such as teaching-induced and developmental factors.

Paquot's (2010) large-scale research on 10 ICLE varieties is a prime example of a CIA-motivated analysis. Common problem areas such as a limited lexical repertoire, lack of register awareness, semantic misuse, and a marked preference for sentence-initial position of connectors, were found across all learner populations.

6.6.5 Corpora in second language acquisition (SLA) research

Granger (2002) has advocated the use of learner corpus data to complement SLA research, which, as she notes, has traditionally drawn on three major categories of data: language use elicited through experimental data carefully controlled for variables, self-report data from introspection and data based on metalingual judgements (Ellis 1994). However, as Granger also points out, much current SLA research tends to be dismissive of naturally occurring language data on account of controlling the variables affecting learner output in non-experimental conditions (see Quote 6.11).

At this juncture, it would be apposite to review another key criticism by SLA specialists against L1 and L2 comparisons. Bley-Vroman (1983) has criticised L1/L2 comparisons for being guilty of the 'comparative fallacy'. He takes issue with those studies comparing learner language to a native-speaker norm as they fail to analyse interlanguage in its own right. This position has been rebutted by Granger (2009), whose essential argument is that a native speaker norm is still present, albeit implicitly, in non-corpus-based SLA studies.

Example 6.20 Possible influence of L1 on learner writing

Gilquin and Paquot (2008) note the overuse of *let's/let us* in French learners' essays which they note may well be the result of L1 transfer as imperatives occur more frequently in written French, and are used with a range of different verbs to structure discourse, as shown by the following examples from French editorials and the accompanying chart (Figure 6.3).

Prenons l'exemple des sorciers ou des magiciens au Moyen Age.
'Let us take the example of wizards or magicians in the Middle Ages'.

Envisageons tout d'abord la question économique.
'Let us first consider economic issues'.

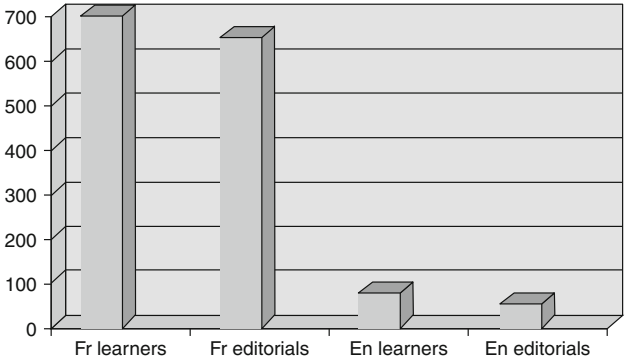

Figure 6.3 First imperative plural verbs in French and English (relative frequency per million words)

Quote 6.11 Granger's rebuttal of the 'comparative fallacy'

It is perfectly possible to do learner corpus analysis without comparing L1 and L2: learner language can be analysed in its own right either cross-sectionally or longitudinally. … It should be pointed out, however, that L1–L2 comparisons are extremely powerful heuristic tools which help bring to light features of learner language which have not been focused on before and which, once uncovered, can be analysed from a strictly L2 perspective. In addition, a look at the non-corpus-based SLA literature shows that the comparative fallacy is in fact pervasive but in a hidden undercover way, to the point that the term 'comparative hypocrisy' comes to mind. For example, all the studies that compare learners of different proficiency levels

are in fact based on an underlying L1 norm as proficiency is usually assessed with an L1 target in mind. The same can be said of SLA studies reporting the results of grammaticality judgment tests. I therefore agree with Sung Park (2004) that 'any SLA study implicitly has a built-in notion of interlanguage with the target language lurking in the background'. The difference is that in traditional SLA studies, the native speaker norm is often implicit and intuition-based while in learner corpus research, it is explicit and corpus-based (Mukherjee 2005).

(Granger 2009: 18–19)

While Granger (1998c, 2002, 2004) has reiterated the value of learner corpora for SLA research, it is surprising that there has not been more of an uptake of her calls for longitudinal studies to complement those taking a cross-sectional dimension. One reason for the paucity of studies in this area may well be due to the difficulty of collecting large quantities of writing across different years and levels (but see Wen et al. 2005 for one such longitudinal study).

Other corpus-inspired SLA studies are those by Tono (2000) on Japanese L1 learners and Housen (2002) on French and Dutch L1 learners, in which learner data are compared against Dulay and Burt's (1975) order for acquisition of grammatical morphemes and of tense/aspect, respectively. Leńko-Szymańska (2002) has investigated the growth in advanced learners' active vocabulary using texts from the PELCRA learner corpus written by first- and fourth-year students of English at the University of Łódź, Poland. As for other languages Belz (2004) examined the use of the *da*-compound by learners of German in a corpus of telecollaborative correspondence, and Collentine and Asención-Delaney (2010) the discourse functions of *ser/estar* + adjective across three levels of Spanish learners.

With the exception of Housen's study, all the others cited above make use of written corpus data. However, Myles and Mitchell (2004) and Myles (2005) make a case for using oral data in preference to written data for the study of SLA.

Quote 6.12 The case for using oral learner data for SLA research

For the purposes of fundamental SLA research, should datasets comprise spoken or written data? To us it seems that oral data must be a better window into the learner's underlying interlanguage system than written data, which may be complicated by extra layers of 'learned' knowledge and monitoring processes.

(Myles and Mitchell 2004: 171)

An SLA research project initiated by Myles (2005) is the French Learner Language Oral Corpora (FLLOC) project, which is modelled on the Child Language Data Exchange (CHILDES) system, (as is Housen's study), a database set up to track first language development across different age groups. Myles (2005) advocates use of the CHILDES system on several accounts, including its data sharing function and most importantly its morphosyntactic tagger and sophisticated search facilities (see Lu 2010 for an overview of this database).

6.6.6 Concluding remarks

Academic genres have been the main focus of learner corpus research. Learner corpora of professional communication such as the Indianapolis Business Learner Corpus (IBLC) consisting of 200 letters of job applications (Upton and Connor 2001) are a rarity (see Case Study 8.2). Moreover, cross-sectional studies, such as those described in the first part of this section, still dominate the field in spite of calls by Granger for more SLA-oriented longitudinal studies. Within the longitudinal studies two somewhat different directions seem to be emerging: studies of written corpora tend to be conducted by researchers coming from corpus linguistics, 'SLA-minded corpus linguists', whereas studies of spoken corpora are generally initiated by those researchers from a background in SLA, 'corpus linguistically-minded SLA researchers'. To what extent these two strands will converge in future remains to be seen.

6.7 Corpora for lexicographic purposes

Without doubt, corpora are now all-pervasive in lexicographic research for practice, i.e. as a resource for the compilation of different types of dictionaries (cf. Hartmann 2001). The use of corpora as a source for lexicographic work has numerous advantages (see Atkins et al. 1994 for a historic, and Heid 2009 for a current, overview of the field).

In the past, lexicographers relied on citation slips, i.e. lexicographic archives, manually extracted from authentic data, a method well chronicled in the development of the *Oxford English Dictionary* (1989). However, as Atkins and Rundell (2008) point out, this method of data collection has several disadvantages. It is exceedingly labour-intensive, resulting in a much lower volume of evidence than can be obtained from corpora. Atkins and Rundell (ibid.) also note that despite the fact that the usage examples are drawn from naturally occurring data, a degree of subjectivity may enter into the selection as humans may select what is interesting rather than what is typical or specific usage. The use of corpora, providing a huge amount of objective data, can overcome both these problems and can also offer other advantages, most notably the access corpora provide to context.

The following sections will review the use of corpora as a resource for lexicography and discuss issues in creating, annotating and utilising corpus data, with particular focus on EFL dictionaries. Much of the discussion will centre on Cobuild (Sinclair 1987), as this was the first corpus-based dictionary of its kind, which laid the foundations for the groundswell of other corpus-based dictionaries from the major publishing houses.

6.7.1 Corpus compilation

In spite of the above advantages afforded by corpora, there are still important issues to consider, especially regarding size, representativeness, balance and content (see Krishnamurthy 1987 for a detailed list of criteria to be addressed at the compilation stage). According to Sinclair (2003a), the size of a lexicographic corpus must be adequate and the content sufficiently heterogeneous to capture *recurrent* language events manifest in different texts by different authors. Nevertheless, rare items pose difficulty for lexicographers as they do not provide enough data on which to adduce evidence and cite authentic examples for dictionary entries. As a case in point, McCarthy (2008) mentions that to obtain sufficient examples of *raining cats and dogs*, a corpus of at least half a billion words would be necessary.

As well as reiterating that size is an important issue in corpus compilation, Atkins and Rundell (2008) also stress that balance and representativeness are other crucial considerations with the proviso that there is no ideal construct for these.

Quote 6.13 Atkins and Rundell on corpus design

> … a lexicographic corpus should be as large and diverse as possible, and … the technical constraints which once made these objectives so challenging have to a large extent disappeared. A truly representative corpus is an impossible goal because we are sampling from a population whose nature is unknowable and whose extent is unlimited. Nevertheless, we know that the description of language in a dictionary cannot be complete if the dictionary's source data doesn't reflect the full repertoire of language events. Our goal, then, is a 'balanced' corpus, though we recognize that there is no single, scientific methodology for achieving this. Texts can be categorized in a variety of ways, but even the very broad categories have fuzzy boundaries and are not always mutually exclusive.
>
> (Atkins and Rundell 2008: 74–5)

The 7-million-word first version of the Cobuild corpus was weighted towards fiction, giving too much prominence to the kind of speech act verbs found in novels. Other early lexicographic corpora included a large amount of

journalistic sources. This explains the inclusion of colloquial expressions in corpus-based EFL dictionaries, as noted by Krishnamurthy (2002: 4) 'EFL dictionaries began to base their headword lists on corpus frequency, and therefore included many more journalistic and colloquial expressions (e.g. OALD6's new words: *cardboard city, generation X, latchkey child, multiskilling, outsource*, *innit*), leaving less space to accommodate literary and higher-register items'.

With the increasing popularity of computer-mediated communication (CMC) such as blogs, e-mail, chat rooms, etc. (see Section 4.6), it is envisaged that dictionaries will also incorporate the type of hybrid language found in this newly emerging electronic register. In fact, the 2-billion-word *Oxford English Corpus* includes such language, but this also raises other questions concerning the integrity of the data, as the texts may not have passed through an editorial process and may be written by non-native speakers. The compilers of the *Oxford English Corpus* have also included language not only from the UK and the United States, but also from Australia, the Caribbean, Canada, India, Singapore and South Africa, subscribing to the view of English as a global language (see Section 5.3.2). But to what extent, if any, dictionaries should represent non-standard varieties of English, such as the eight different varieties in the *Collins Wordbanks Online Corpus* (or, indeed, varieties of other languages such as Canadian French), is another moot point.

Lexicographic corpus compilation could thus be viewed as fraught with difficult choices regarding sources, representativeness and balance, considerations for which there exist no ideal solutions, but rather approximations towards achieving these. As Béjoint (2010) states, these issues have hardly changed throughout the years. It is only with the advent of lexicographic corpus investigations that the challenges besetting traditional lexicographers have come to the fore and been articulated so clearly and extensively.

6.7.2 Corpus annotation

There is a wealth of publications on corpus annotation, the most well known of which is probably the CLAWS tagging system developed at Lancaster University (cf. Garside et al. 1997). Annotation mainly involves tagging and parsing of a corpus.

Concept 6.15 Corpus annotation

Part-of-speech (POS) tagging involves the automatic assignation of a word class to every word in the corpus. Some tagging systems are quite sophisticated; for example, CLAWS has 57 grammatical tags, while the BNC has different tags to distinguish different types of nouns, i.e. singular, plural, proper and common nouns. Parsers, on the other hand, assign syntactic

categories such as subject, verb and object to sentences. However, unlike POS taggers which have 97 per cent accuracy, parsers only have a success rate of about 75 per cent. For example, Atkins and Rundell (ibid.: 92) give the following problematic sentences, which a parser may not be able to cope with, as it cannot distinguish whether the prepositional phrase attaches to *treating* or *patients*.

guidelines for treating patients with AIDS
guidelines for treating patients with antibiotics

Although software tools for annotation of corpora have some obvious advantages, for lexicographic purposes for which a great deal of data is needed as noted above, it may be best to sacrifice careful, time-consuming annotation for a more 'quick and dirty' approach. Atkins and Rundell (ibid.) argue that lexicographers are interested in regularities and norms in language and that a few imperfect results from annotation would not really matter, as a general rule advocating size over granularity.

One criticism levelled by Sinclair (2003b) against annotation concerns their paradigmatic orientation as annotation schemes bypass the syntagmatic nature of language. Nevertheless, Sinclair does not deny that statistical tools have their merits; they just have to be used judiciously. Lexical profiling used on a POS-tagged corpus is a very useful approach for lexicographic work as it can accommodate the syntagmatic nature of language (see Case Study 8.4).

Concept 6.16 Lexical profiling

Lexical profiling is an efficient way to exploit a large corpus while reducing the effort required by the human user. Word Sketches (cf. Kilgariff and Rundell 2002) produces a kind of statistical summary which reveals the grammatical and collocational behaviour of a word. For example, a Word Sketch for the verb *exercise* reveals that the kinds of objects *exercise* usually takes are words like *restraint, discretion, caution*, and *vigilance* (Atkins and Rundell ibid.: 91).

6.7.3 Corpus utilisation

It is the insights afforded by the 'virtually unlimited context' in which concordance lines can be viewed (Čermák 2003) that have reconceptualised the ways in which word senses, definitions, the relationship between meaning and form, and grammatical information are encoded in the dictionary.

Béjoint (2010) refers to the notion of lexicographers being either 'lumpers' or 'splitters', citing Hanks's (2002: 159) observation that a word in a dictionary 'can have about as many senses as the lexicographer cares to perceive'. Lexicographers working in the corpus tradition tend to be splitters rather than lumpers such that many more senses are now recorded, with COBUILD the first dictionary to include *see* in the sense of understand.

The examination of lexical items within their naturally occurring wider context has helped to elucidate the link between meaning and form, disambiguate the different senses of a word and give a deeper insight into polysemy, homonymy and metaphor (see Moon 1987). By way of illustration for the role of context in disambiguation of word senses, Atkins and Rundell (ibid.) take the case of *issue*.

Example 6.21 The role of context for disambiguation

Atkins and Rundell (2008: 295) note that corpus data show *issue* to have acquired a new sense recently, i.e. personal problems or difficulties, as evidenced by the following:

> Due to his emotionally chaotic upbringing, he likely does have significant intimacy issues.
> It helps to understand … how issues around gender, dependency, daily routine and staff responsibility impinge on the environment.

Atkins and Rundell examine the factors that distinguish this new use of *issue* from its basic sense. First, this usage tends to be found in areas relating to social science, e.g. psychology, counselling, criminology and childcare. Linguistic features of a syntagmatic nature which set it apart from the basic sense are the following:

- It is always (in this use) pluralised.
- It is rarely sentence-initial or clause-initial, but usually occurs in the patterns *have* + *issues* or *with* + *issues*.
- It is often followed by *around* (with the meaning of 'concerning')
- It is often premodified by another noun (as in *intimacy issues* or *incontinence issues*).

The previous examples all testify to the greater attention paid to the syntagmatic nature of language, a feature that was incorporated into the Cobuild system for recording of grammatical notes: 'The system of syntax notation was specially developed to allow lexicographers to record not only word classes and

paradigmatic syntactic variations, but also individual syntagmatic sequences' (Krisnamurthy 1987: 66). The gulf between the pre-corpus learners' dictionaries and corpus-based ones with respect to the paradigmatic/syntagmatic axis has been commented on by Hoey and Brook O'Donnell (2008).

Example 6.22 Pre-corpus vs corpus-based dictionaries

A comparison of a pre-corpus learners' dictionary such as OALD3 (1974) with Cobuild1 (1987) reveals some important differences. Consider the use of *move* as a noun. This reveals a shift in grammatical stance between the two dictionaries. OALD was written at a time when the grammar/lexis dichotomy was unquestioned: any use of *move* as a count noun in a grammatical structure permitting the use of a count noun should in principle be acceptable, subject to certain constraints. Cobuild, however, pays more attention to detailed grammatical norms – for example it notes that in phrases such as '*make a move*' the noun colligates paradigmatically with the singular and syntagmatically with the verb *make*; in the expression '*get a move on*' it colligates syntagmatically with a verb and an adjunct.

Many noun uses of this word are closely linked semantically to a basic verb sense. Cobuild represents these by appending '▸used as a noun' to the relevant verb sense, showing thereby that the word has a secondary colligation with the noun class.

(Hoey and Brook O'Donnell 2008: 294)

Findings from learner corpora (see Section 6.6) are now being increasingly utilised in the compilation of learners' dictionaries. Gillard and Gadsby (1998) illustrate how learner errors in the 10-million-word Longman Learners' Corpus (LLC) have been used to help compile the *Longman Essential Activator* (*LEA*) (1997). Paquot's (2010) extensive research findings on expert and learner corpora of academic writing have been used as input for 'help' boxes in the *Macmillan English Dictionary* (Rundell 2002).

6.7.4 Dilemmas for lexicographers

While lexicographic corpora have revealed a wealth of valuable information on frequency counts and context, at the same time this additional information has created dilemmas for lexicographers. As regards the different senses of a word, the abstract/metaphoric sense may be more frequent than its concrete sense. But this does not necessarily entail that its most frequent sense occupies the first entry in the dictionary. J. Flowerdew (2009) points out that in the COBUILD dictionary 'teachability' seems to have been the principle adopted for listing of entries, with concrete meanings superseding abstract ones; a case

in point is the concrete meaning of 'lifebelt' coming before its metaphorical sense, although it is the metaphorical sense which is the more frequent (see Section 7.2.1).

Another dilemma for lexicographers is how much of the syntagmatic sequences uncovered through the 'virtually unlimited context' can, or should be, encoded in the dictionary. While one can argue that with restrictions on space eased, online dictionaries will be able to incorporate links to a wealth of linguistic phraseological data hitherto neglected, this is only part of the answer. Stubbs (2009: 131), while calling for the compilation of a 'phraseological dictionary', foresees difficulties in achieving this on account of the probabilistic nature and fuzzy boundaries of phraseological sequences.

Quote 6.14 Stubbs on the problems of a 'phraseological dictionary'

> The major descriptive problem is that the units are internally variable and have indeterminate boundaries, and that similar units are often related to each other in taxonomic hierarchies (Croft 2001: 25; Stubbs 2007). The general solution is to describe their canonical, prototypical forms, but deciding on the appropriate level of delicacy for different purposes is a matter of interpretation.
>
> (Stubbs 2009: 131)

The main challenges for future corpus-based lexicography would thus seem to lie in how to represent these more discourse-based aspects of phraseology, such as semantic prosody, which is often realised at a level above the sentence (cf. Stubbs 2001a).

6.7.5 Concluding remarks and future directions

While the focus of discussion in this chapter has mainly been devoted to how the COBUILD corpus has been used to inform the design of EFL dictionaries, it should not be forgotten that the 100-million-word BNC has also played a role in dictionary compilation, but of a different sort. For example, the BNC provided input for the *Oxford Collocations Dictionary* (2002) and the *New Oxford Dictionary of English* (1998) aimed at native speakers rather than EFL learners. The BNC and a corpus of French were used as input for the *Oxford-Hachette French Dictionary* (1994). Corpus-based innovations are slowly filtering through to bilingual dictionaries with the increasing availability of multilingual corpora and research thereof (Heid 2009).

Notwithstanding the difficulties in capturing phraseological information in an online dictionary, as noted above, a few initiatives are underway to

accomplish this. Hanks (2009) outlines the basis of *The Pattern Dictionary*, which provides explicit links between meaning and use, and aims to capture the phraseology of all the patterns of a particular verb. Another initiative underway is the FrameNet project (Baker et al. 2003), which takes a different conceptual starting point by grouping words with similar meanings into semantic frames. The online dictionaries of the future with their focus on the phraseological aspects of language will thus bring about a confluence of traditional grammar and dictionary information (see Section 7.2.1).

6.8 Corpora for testing purposes

6.8.1 Enhancement of language testing and assessment materials

In a 1996 article Alderson asked whether corpora have a role to play in language testing – this section aims to discuss whether corpora have fulfilled their initial promise in the intervening years. Corpora can enhance language testing by being used in the following ways:

- to archive examination scripts
- to develop test materials
- to improve the quality of test construction
- to validate tests
- to standardise tests

A related application would be for test coaches to ensure quality control. Other aspects to consider are which types of corpora are suitable to meet the above aims and whether they are being used in the context of large-scale, high-stakes standardised assessment or in local institutional settings for less formal, low stakes approaches to assessment. The following sections will exemplify how corpora have brought similar advantages to language testing as they have to language teaching (see Chapter 7), but their focus is somewhat different. Whereas most of the initiatives in the application of corpora to language teaching are at a local institutional level, those endeavours in applying corpora to testing and assessment have been carried out at national or international levels. Another difference is that although learner corpora are still on the periphery of language teaching, they are at the core of language testing initiatives, which are discussed below from the perspective of meeting specific testing and assessment situations.

6.8.2 Applications at the international level

The University of Cambridge Local Examination Syndicate (UCLES) and the Educational Testing Service (ETS) in the US have both set up large-scale projects

for the compilation of corpora for various purposes in language testing. Their major initiatives are outlined below.

TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL)

The T2K-SWAL Corpus was constructed, with support funding from the ETS, to investigate university level language skills required by candidates taking the TOEFL test. The corpus is made up of 2.7 million words from 423 written/spoken texts. Not only did Biber et al. (2004) compile corpora of the traditional academic registers, i.e. research papers and academic lectures, but they extended their analysis to include other registers such as handbooks, Web pages, service encounters and classroom management talk. This was done in response to the lack of knowledge about the linguistic characteristics of these academic registers in order to provide a more representative basis for test construction and validation, specifically the TOEFL test.

Quote 6.15 Rationale for T2K-SWAL Corpus

> Given our lack of basic knowledge, it has been nearly impossible to evaluate the extent to which textual materials for ESL/EFL instruction and assessment actually represent the linguistic characteristics of the target registers. The one directly relevant study that we know of, Biber and Jamieson (1998), indicates that TOEFL exam texts are often quite different from the target registers. For example, passive constructions were much more likely to occur frequently in long conversations and lectures from TOEFL exams than in the associated target registers. Further, there was an extremely wide range of variation within the categories of *long conversations, lectures* and *reading passages* with respect to their use of relative clause constructions and passive constructions.
>
> (Biber et al. 2004: 2)

A key motivation of the project was to provide an empirical set of data to replace the intuitions of writers in constructing tests and test compilers in deciding what to test. In this connection, the findings of a research project conducted by Alderson (2007) are salutary. Eighteen judges who were linguists working at Lancaster University completed three separate tasks in which they were asked to make judgements on frequency counts; for example, one of the tasks demanded that they examine a list of 100 verbs to decide how frequently each verb occurred per million words of English. All three investigations of frequency judgments showed that the judgments made by professional linguists did not correlate highly with corpus-based frequency counts, with Alderson concluding that 'further research is needed into the nature of intuitions about word frequency' (p. 383).

Cambridge Learner Corpus (CLC)

The T2K-SWAL corpus discussed above has been designed for one specific test, i.e. TOEFL, with a focus on academic registers. The CLC, on the other hand, is primarily an archive 'to be used for both general and specific test development and validation projects across many different types and forms of test' (Taylor and Barker 2008: 249). In fact, Barker (2008a) notes that corpora did not really make a significant impact on language testing and assessment until the development of learner corpora. An overview of this archive, both the spoken and written components, is given below (see also Barker 2010).

Concept 6.17 Overview of Cambridge Learner Corpus

Cambridge Learner Corpus
30 million words
Range of exams: Main Suite, BEC, CELS, IELTS
Error-coding of 50 per cent of scripts (see Nicholls 2003)
Most Common European Framework of Reference (CEFR) levels (A2–C2)
Linked demographic/score data
Wide L1/nationality coverage (91 source languages)
Wide proficiency coverage

Cambridge Spoken Learner Corpus
8500 speaking tests
All Cambridge ESOL exams (CEFR levels A1–C2)
Range of types of English: general academic, business and young learner English

(Barker 2008b)

The Cambridge tests correspond to the six different levels of the Common European Framework of Reference for Languages (CEFR), as described below.

Quote 6.16 Common European Framework of Reference for Languages (CEFR)

One of the principal aims of the *Common European Framework of Reference for Languages* (CEFR; Council of Europe 2001) is to function as a kind of common language, a 'lingua franca' in the modern languages field that allows language courses, curricula, syllabuses, teaching and learning materials, tests and assessment systems to refer back to a common scale of measurement, thus enhancing comparability and transparency.

(Salamoura 2008: 5)

Small-scale exploratory studies have been carried out to provide insights into item writer training (Hargreaves 2000; Rose 2008), to define the construct of reading and validate reading tests (Barker 2008b), all using comparative data from the BNC. A larger-scale investigation, making use of both intuitive and corpus-assisted approaches, was also carried out on key language features in writing scripts across four bands/levels to distinguish performance in writing with suggestions for how these could be incorporated into a common scale for writing (Hawkey and Barker 2004), although no mention is made of any contradictions that might have arisen between testers' intuitions and corpus-based evidence, as in the Alderson study.

Example 6.23 Integrating manual and computerised analyses

In Hawkey and Barker's (2004) study, from a close reading of the scripts a draft working scale began to emerge using three criteria: *sophistication of language*, *accuracy* and *organisation and coherence*. Following the manual analysis, WordSmith Tools was used to investigate:

- whole script, sentence and paragraph lengths
- title use
- vocabulary range
- words in concordances and collocations; and
- errors

The computer analysis of the *range of vocabulary* of each subcorpus, i.e. band, was found to support inferences from the manual analysis, and could be regarded as a feature distinguishing proficiency levels.

Band	*Tokens*	*Types*	*Type: token ratio**	*No. of scripts*	*Av. types per script*
5	6130	1191	19.43	29	41
4	3112	619	19.89	18	34
3	7999	1116	13.95	43	26
2	1206	342	28.36	8	43

* The type: token ratio column expresses the number of different words in each subcorpus as a percentage of the total number of words in that subcorpus.

The type token ratios were found to be higher at Bands 4 and 5, indicating that more lexical items are used at these proficiency levels. Hawkey and Barker (ibid.) explain the high type : token ratio for the Band 2 scripts by the low number of scripts in this sub-corpus. The size of a corpus is a perennial problem in corpus work to which there are no easy answers as size is dependent on so many variables, e.g. how conventionalised the genre is, the items under investigation, etc. (see Section 1.1).

(Adapted from Hawkey and Barker 2004: 152)

6.8.3 Applications at the national/institutional level

Apart from the two large-scale initiatives described above, there also exist several other testing and assessment projects. One project related to CEFR is that by Osborne (2011). Osborne's study focuses on how the findings from a learner corpus of spoken English, the PAROLE Corpus, can provide evidence for fluency features at different CEFR levels and thus be used for benchmarking samples of oral production.

Quote 6.17 Description and aims of the PAROLE Corpus

> The *PAROLE* corpus (Parallèle, Oral, en Langue Etrangère) consists of 15–20 minute recordings of speakers of L2 English (L1 French and German), of L2 French (various L1s) and of L2 Italian (L1 French), along with recordings from native speakers of these three languages carrying out the same tasks. The recordings are … annotated for pauses (filled and unfilled), retracings and errors, with a view to comparing (dis)fluency characteristics across languages, between native speakers and non-natives, and between non-native speakers at different levels of proficiency.
>
> (Osborne 2011: 181)

Osborne's study included both temporal measures of fluency (speech rate, pauses, length of utterances, retracing) and more qualitative aspects such as propositional content, with an additional measure of 'granularity', i.e. how much detail a speaker provides, syntactic density, vocabulary range and accuracy. The results of this study so far indicate that 'independent CEF ratings correlate best with speech rate (wpm) and percentage of pause time, and least well with retracing and granularity. For any individual speaker, a single measure taken in isolation is not necessarily a reliable indicator of proficiency, so that overall fluency is best measured as a group of features' (Osborne 2011: 185). Another testing-related project is that by Hasselgren (2002) who examined smallwords, e.g. 'well' and 'sort of', as markers of fluency, which Osborne's study did not seem to specifically target. Through corpus analyses of learner and native speaker English she demonstrated that smallwords created coherence, connecting and organising text, concluding that such evidence could lead to the establishment of descriptors of fluency involving fairer judgments.

It was noted in the previous section that 50 per cent of the scripts in the Cambridge Learner Corpus have been error-tagged (Nicholls 2003), which may be of help in designing corpus-based language tests. Coniam (1997b), for example, describes how a multiple choice vocabulary cloze test can be produced from a text, which involves assigning word class tags to the text and then retrieving word frequencies for the words in the text from an analysed corpus, in this case the Bank of English. Coniam maintains that this operational

principle does have a certain amount of validity as several previous studies on vocabulary have shown there to be a relationship between word frequency and proficiency levels. Although the program does have several drawbacks, for example it cannot distinguish between different senses of a word for the same word class, nevertheless it is distinctive for being one of the first forays of its kind into corpus-based language testing.

6.8.4 Future prospects

The main thrust of UCLES' ongoing research is to draw up reference level descriptions for English which are aligned with the six levels of the CEFR, but which expand on and exemplify the existing CEFR scales. To this end, a new learner corpus is being built, the samples of which will consist of language 'produced by learners *on demand* and *for* the corpus' (Alexopoulou 2008: 15), similar to the tasks in the ICLE corpus. Ädel (2006), in her analysis of metadiscourse in L1 compared with that of L2 English in the Swedish component of the ICLE corpus, notes the effect of task, specifically untimed essays, on learner performance. What are now needed are studies which examine task effect on performance testing. Another key feature of this new corpus is that more data will be collected relating to the learners such as their age and the educational and sociocultural settings in which they operate, thus reflecting the more sociolinguistic approach the field of corpus linguistics seems to be taking in general (see Chapter 5).

One area targeted for future development by UCLES concerns the compilation of domain-specific corpora (see Wright 2008). Although such corpora are now well developed for ESAP, their application to language testing and assessment has yet to be fully realised. Moreover, Taylor and Barker (2008) have brought up the issue of comparing test-taker responses with native speaker corpora, questioning the usefulness of this tradition in light of the 'English as a lingua franca' debate (see Section 6.1). Such concerns presage the compilation of corpora for other language varieties as a consideration in language testing and assessment.

This broadening of the field in using corpora for testing purposes is also in evidence in research being undertaken by ETS. Studies are being carried out which are looking not only at the final essay product but also the process of writing in light of the observation that 'we suspect that there are limits to the amount of information that can be wrung from a student essay without additional sources of data' (Deane and Quinlan 2010: 172). Detailed profiles of students' prewriting, drafting and revising are being captured in an attempt to understand the underlying cognitive processes in order to better identify profiles of writing behaviour to enhance test construction.

Except for the two large-scale projects initiated by ETS in the US and UCLES in the UK, the use of corpora for language testing and assessment purposes

still seems to be in its infancy as far as applications in more local settings are concerned. Granger (2004), cited in Taylor and Barker (2008), distinguishes between commercial and academic corpora. The fact that collecting, transcribing and analysing corpora is an extremely expensive undertaking often requiring commercial funding may no doubt explain the scarcity of small, specialised corpora being used at the local level. However, with increasing interest in this area, development of software tools, and hopefully more publicly available corpora, the future will see the compilation of more locally compiled corpora to redress this imbalance.

Further reading

O'Keeffe, A. and McCarthy, M. (eds) *The Routledge Handbook of Corpus Linguistics*. Section VII, 'Using corpora to study literature and translation', covers key themes in these fields.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge. Through the use of research questions, case studies and discussion of methodological issues, this volume presents a comprehensive overview of the area.

Prodromou, L. (2008) *English as a Lingua Franca: a Corpus-Based Analysis*. London: Continuum. This volume reviews key debates in the field of ELF.

Toolan, M. (2010) *Narrative Progression in the Short Story: a Corpus Stylistic Approach*. Amsterdam: John Benjamins. This volume investigates narrative prospecting, i.e. how a text guides reader judgements, through a variety of corpus analytic methods and tools including *WordSmith Tools* (Scott 2004) and *WMatrix* (Rayson 2005).

7
Applying Corpus Linguistics in Teaching Arenas

This chapter will:

- Discuss key issues relating to the pedagogic relevance of corpora
- Discuss direct and indirect pedagogic applications of corpora
- Present the main impediments to data-driven learning (DDL)
- Profile under-represented corpora in DDL
- Present initiatives in using corpora for teaching content courses

In this chapter, Sections 7.1–7.4 focus on the application of corpora in the teaching of skills-based courses. Sections 7.5–7.8 examine the use of corpora in content-based teaching areas, namely translation, teacher education, and literature.

7.1 The pedagogic relevance of corpora: some key issues

One key issue discussed in the literature concerns to what extent corpus findings and statistical data can, or should, have a bearing on teaching materials. Cook (1998) distinguishes between 'hard' and 'soft' approaches regarding the relevance of corpus findings to language teaching. Taking the soft line view, he advocates that pedagogic materials should be 'corpus-based and not 'corpus-bound', viewing corpora as a *contribution* to, rather than a solid base for materials.

Quote 7.1 Cook on applications of corpora to pedagogy

> But where pedagogy is concerned, corpus statistics say nothing about immeasurable but crucial factors such as students' and teachers' attitudes and expectations, the personal relationships between them, their own wishes, or the diversity of traditions from which they come. Consequently,

> computer corpora – while impressive and interesting records of certain aspects of language use – can never be more than a *contribution* to our understanding of effective language teaching.
>
> (Cook 1998: 58)

On the question of corpus statistics, both Widdowson (1991) and Cook (1998) underscore the danger of equating frequency with pedagogic relevance. Indeed, other factors come into play such as the utility value of items: an item may be infrequent but salient (see discussion of salience in Section 1.2.1), or 'frequent but limited in range, or infrequent but useful in a wide range of contexts' (Cook 1998: 62). Ishii (2009) makes the point that certain verbs found in phrasal verb dictionaries such as 'catch up with' should still be taught for their salience to learners in spite of their low-frequency data in general-purpose corpora, such as the BNC.

Quote 7.2 Widdowson on criteria for pedagogic selection

> … words and structures might be identified as 'pedagogically' core or nuclear, and preferred as a prototype at a particular learning stage because of their coverage or their generative value, because they are catalysts which activate the learning process, whatever their status might be in respect to their actual occurrence in context of use.
>
> (Widdowson 2003: 87)

The pedagogic selection of materials also relates to whether they are for productive or receptive use, but this is very rarely touched on in the literature. Rebutting Cook's (1998: 61) argument on the drawback of using authentic spoken corpus data as a great deal of it can be 'inarticulate, impoverished and inexpressive', Campbell et al. (2007) point out that this standpoint needs to be considered within the context of whether the authentic input data is being used for listening purposes or presented to the learner for emulation. If for productive purposes, then there may be a case, as McCarthy (1998) has argued, for mediating the corpus in order to filter out unnecessary 'noise'. On the other hand, a case can be made, as Campbell et al. have argued with respect to their Chinese students studying in Dublin, that exposing learners to samples of 'real' spoken language as found within the Dublin community will help learners integrate more easily into their present language community and understand the sociocultural environment of that target community.

In their defence for using a corpus of authentic speech, Campbell et al. (ibid.) revisit Widdowson's (2000) objection to corpus data being considered 'real' language as it has been removed from its original habitat and needs to be 'authenticated' to suit the students' own context (see Mishan 2004; Kaltenböck and Mehlmauer-Larcher 2005; Seidlhofer 2003 for a review of this issue). While Campbell et al. (ibid.) acknowledge the validity of Widdowson's claim, they put forward the counter-argument of the educational value and relevance of such data for *observation* of language, i.e. for receptive purposes, rather than that of a *participant*, i.e. for productive purposes. If used for written purposes, then, as Widdowson has pointed out, corpus data may well need to be authenticated to suit the students' own context of writing (see L. Flowerdew 2009).

Corpus-based research on different academic genres has also reignited the debate on the 'common-core hypothesis' (Bloor and Bloor 1986). Proponents of this approach maintain that there exists a common set of linguistic structures and vocabulary that will be found across a range of academic texts, regardless of discipline and genre. Coxhead's (2000) corpus-based research of vocabulary items in a 3.5-million-word corpus composed of written academic texts drawn from the disciplines of arts, commerce, law and science, seems to support the hypothesis with its extraction of common vocabulary items forming an academic word list (AWL). However, Hyland and Tse's (2007) corpus data point to a discipline-based specific lexical repertoire. Other ESP studies of research articles in which wordlists from specialist disciplines are compared with Coxhead's academic word list also favour sets of discipline-specific lexis (see L. Flowerdew 2011a for a summary of these). It is to be noted that all the studies arguing for a disciplinary specific lexis for EAP are based solely on frequency counts of some kind. Related studies by Ellis et al. (2008) and Simpson-Vlach and Ellis (2010), besides frequency information, also take into account pedagogic relevance as seen through teachers' eyes, and psycholinguistic salience, to arrive at a pedagogic list of formulaic sequences for teaching academic speech and writing. This triangulated study supports Coxhead's AWL (as does Paquot's (2010) research) thus lending weight to the common-core hypothesis, and also echoing the sentiments expressed by Cook and Widdowson that pedagogic relevance may not be circumscribed solely by frequencies found in a corpus.

7.2 Pedagogical corpus applications: indirect and direct

In spite of some doubts expressed about the pedagogic relevance of corpora and issues related to frequency lists, corpora have been extremely influential in informing various aspects of pedagogy over the last few years. These applications can be viewed, in broad terms, as indirect and direct applications (Figure 7.1 as Concept 7.1).

Concept 7.1

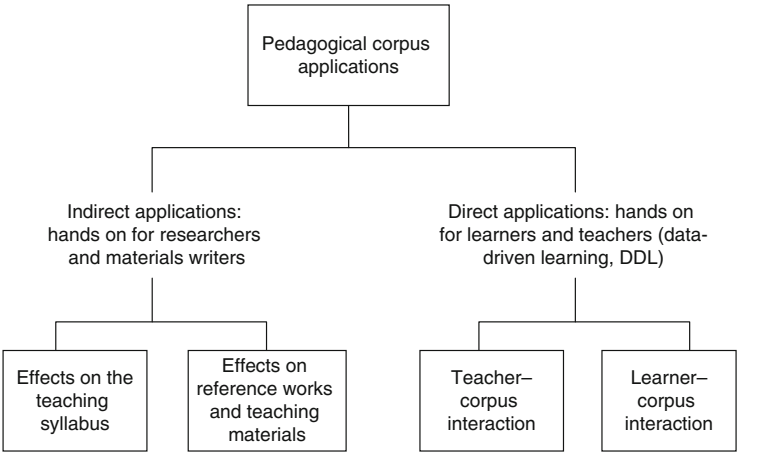

Figure 7.1 Types of pedagogical corpus applications (Römer 2010: 19)

7.2.1 Indirect applications

To inform reference materials

Section 6.7 has discussed how research findings from lexicographic corpora have provided the phraseological, theoretical underpinning for corpus-based dictionaries. Corpus-based grammars have also been constructed along similar philosophical lines. Although these grammars are a remarkable achievement, their drawback for EFL learners is that there may be omissions on account of various factors.

Size may be one limitation. For example, Tucker (2001) carried out a microscopic corpus analysis of *possibly* using the 450-million-word Bank of English corpus to compare its grammatical environments with the grammar notes provided in the *Longman Grammar of Spoken and Written English* (LGSWE), compiled from frequency data in a 40-million-word corpus spanning four different registers of English (conversation, news, academic prose and fiction). Tucker found that the LGSWE omits any explicit reference to the use of *possibly* in non-clausal environments, e.g. 'Mr Murdoch will have to sell assets, possibly at distress prices' (p. 186).

The actual composition of the corpus may be another limitation. For example, the groundbreaking *Collins COBUILD English Grammar* (Sinclair 1990) was initially based on a 20-million-word corpus containing quite a substantial

amount of journalistic text. Meanwhile, the *Cambridge Grammar of English* (Carter and McCarthy 2006) which could be considered as the third-generation corpus-based grammar, draws on the 700-million-word Cambridge International Corpus made up of texts taken from everyday spoken and written English, which, notwithstanding the size of the corpus, may not be fully representative of a range of genres.

On account of the above limitations, these three corpus-based grammars may ignore linguistic facts derived from native-speaker introspection, which can capture subtle interpretations in meaning possibly overlooked in the corpus grammars (see Mukherjee 2006 who suggests combining the two grammars for the best of both worlds). Corpus-based grammars need to improve their comprehensiveness, but how is this to be achieved? In a paper summing up a panel discussion at the 24th ICAME conference, Aarts summarises views as follows, suggesting a merging of traditional grammar and dictionary information, similar to the proposals made for online dictionaries (see Section 6.7.5).

Quote 7.3 Aarts on the internet grammar of the future

> Panel and audience alike were of the opinion that the reference grammar of the twenty-first century will be on the internet: ideally, it will offer information about syntactic structures in combination with lexico-grammatical information about lexis-dependent patterns, but also in relation to information about collocations, multiword units and stereotyped phrases. The information will be layered, in that it offers different levels of access, such as the learner's level and the researcher's level.
>
> (Aarts 2006: 403)

To inform the teaching syllabus and ELT materials

Around the same period as the compilation of the COBUILD dictionary, the same corpus-theoretical approach in which the lexical item has primacy (Sinclair 1999) inspired the lexical syllabus (Sinclair 1998; Willis 1990). This provided the foundation for a set of teaching materials (Willis and Willis 1988).

Concept 7.2 The lexical syllabus

> Instead of being organized in terms of grammatical forms, the syllabus can be designed around the most important recurrent patterns … This type of syllabus is referred to as a lexical syllabus, although this is somewhat

misleading, as it is designed around lexical patterns, not single words. The idea of basing a syllabus on patterns of use, was, in fact, put forward as early as 1980 by Nattinger (1980).

Although the emphasis is on lexical patterning in the lexical syllabus, grammar is not neglected, it can be argued, as the main lexical patterns will incorporate the main grammatical forms. Willis (1990: vi) takes this a stage further claiming that 'the lexical syllabus not only subsumes a structural syllabus, it also indicates how the structures which make up the syllabus should be exemplified'. For the COBUILD course for the first level the most frequent 700 words were selected from the corpus, these words accounting, according to Willis (1990: vi), for around 70% of all English text. The underlying principle of the lexical syllabus is frequency. Sinclair and Renouf (1988) argue that the most frequent words typically have a range of uses and that many of these uses are typically not covered in beginners' courses. They give the example of the word *make*. This word most typically occurs in patterns such as *make decisions, make discoveries, make arrangements*. These abstract uses are more frequent than the concrete use, as in *make a cake*.

(J. Flowerdew 2009: 335)

In spite of a few recent endeavours to produce corpus-based ELT textbooks such as those in McCarthy et al.'s (2005) *Touchstone* series, the majority of ELT textbooks are non-corpus based. A number of studies have compared various linguistic features as they are covered in such ELT textbooks with corpus findings. For example, Yoo (2009) used the corpus findings from the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). Yoo's (ibid.) comparison centred on the findings of the treatment of the English definite article across six ESL/EFL Grammar Series (21 books in total). One key finding of this study was that the corpus showed cataphoric use (postmodification) to be the most common type of definite article usage in academic prose, accounting for 40 per cent of all instances of *the*. However, this specific usage received the least attention of all definite article usage in the ESL/EFL grammars surveyed. Yoo surmises that the complexity of this construction may be why most grammars do not discuss this phenomenon.

As regards the language of textbooks, several practitioners have compiled 'pedagogic corpora', i.e. corpora of ELT textbooks (cf. Römer 2004; Meunier and Gouverneur 2009), to compare the language with a corpus of authentic English in order to check the extent of authenticity of the language presented to learners.

Example 7.1 A comparison of modals in the BNC and a school textbook

Römer (2004: 193) compares the use of modals from the spoken component of the BNC with their use in a textbook, *Green Line*, used for teaching English in secondary schools in Germany. Her comparison reveals large discrepancies between the use of modals in the BNC and in the English taught in German schools:

> For *can* and *could* expressing an ability the percentages in *Green Line* (52.5% and 78.3%) are much higher than in the BNC (36% and 34%). In the sentences from the BNC *could* more frequently expresses a possibility (in 41.5% of the cases) than an ability. Concerning *may* we get a much higher share of the permission meaning in *Green Line* (41.7%) than in spoken English (13%), although the modal is mainly used to convey the meaning 'possibility' in actual language use (83%).

It is perhaps no surprise to see these differences given that the textbook is not corpus-based; no doubt the writer has chosen certain items for inclusion in the syllabus according to similar criteria to those proposed by Widdowson and Cook.

Gouverneur's (2008) comparison of collocations across textbooks in the TeMa corpus (corpus of textbook material) showed a lack of consistency, with very few collocations common to all textbooks. She further makes the point that textbooks tend to focus on high-frequency verbs at the intermediate level, but downplay their importance at higher levels because they are considered to have been mastered, which 'is most regrettable, since, as learner corpus evidence has demonstrated, this is far from being the case' (p. 235).

But to what extent should textbooks be modelled on the language in corpora? In fact, Shortall (2007) argues for the use of contrived sentences in textbooks on the grounds that the textbook writers may want to deliberately exaggerate, for example, the frequency of time adverbials with the present perfect tense to highlight their function as time markers. Through such a strategy, those features that facilitate processing are made salient, which may not be the case with corpus data. This is akin to Cook's (2001) argument for using invented sentences as a means to promote 'noticing' in the sense of having students pay conscious attention to the input. For Cook, whether the sentence is attested corpus data or invented is not the issue; what matters is the pedagogic merit of either type of data and how best to promote the learning of a particular item after it has been noticed.

While publishers could usefully exploit both the findings from pedagogic corpora, native speaker and learner corpora to ensure that textbooks more accurately reflect language as it is actually used, the pedagogic relevance of 'real' language vs 'contrived sentences' requires careful consideration, as Cook (2001) and Shortall (2007) have cautioned.

7.2.2 Direct applications (DDL)

While corpora have made enormous contributions to the compilation of dictionaries and grammars, their influence has been less pronounced as far as direct applications (DDL) are concerned, as noted by Stubbs (2004). Tim Johns' (1988, 1991) pioneering work in DDL cannot be underestimated and has been the foundation and inspiration behind pedagogic applications over the past two decades (see Chambers 2010 for a brief history of DDL).

DDL is usually associated with an inductive, discovery-based approach to learning in which students work out rules or probabilities from the examples provided. However, in reality, much corpus-based work also draws on the deductive approach, as pointed out by Johansson (2009).

Quote 7.4 Inductive vs deductive approaches

> Is the use of corpora to be grouped with the explicit or implicit method? The term 'data-driven' learning suggests that it is an inductive approach and therefore comparable with the implicit method, though the emphasis is on gaining insight rather than establishing habits, and in this sense it is mentalistic. I believe that the dichotomy explicit–implicit is far too simple. In the case of corpora in language teaching, I would favour a guided inductive approach or a combination of an inductive and deductive approach where the elements of explanation and corpus use are tailored to the needs of the student.
>
> (Johansson 2009: 41–2)

At the extreme end of the inductive ↔ deductive cline, we have the notion of the 'learner-as-researcher' proposed by Bernardini (2002, 2004). As Kennedy and Miceli (2010) point out, though, this approach makes high demands on the students in terms of language proficiency, observation and inductive reasoning. As their students are intermediate level Italian and not advanced students of translation like those of Bernardini, they propose a more guided-inductive approach through apprenticeship training (see Section 7.4.1).

Another issue relates to the nature of the corpus activities and whether these should be 'pen-and-paper' based or 'hands-on'. Although some practitioners have expressed misgivings about reworking corpus data on account of the fact

that it no longer represents 'real language', teacher-generated materials do have certain advantages. One advantage of having students work with worksheet output of concordance data is that it is a valuable means of providing them with 'corpus competence', thereby gently familiarising them with corpus methodologies such as the inductive approach, interpretation of frequency data, etc. Another advantage is that it allows teachers to sift through what may be a vast number of concordance lines to reduce and select data on the basis of utility value. However this 'corpus-based' approach has its limitations in that it does not allow students to follow a more learner-centred, 'corpus-driven' approach, allowing them to browse the corpus to search for answers to their own queries.

It has also been suggested that field independent learners who are known to prefer instruction emphasising rules may not like the inductive approach to grammar inherent in DDL; this approach may be more appealing to field dependent students who thrive in cooperative, interactive settings (L. Flowerdew 2008b). A flexible tool such as the Chemnitz Internet Grammar, capitalising on different modes of learning by incorporating a rule-based grammar containing grammatical explanations, examples and exercises, and a corpus with an interactive exercise component, would offer students the flexibility to choose their own grammar pathways (cf. Hahn 2000; Schmied 2006a).

In Section 6.2 the lack of corpora of professional written documentation was noted so it is no surprise to find that there are hardly any corpus applications in this field. While several corpora of workplace interactions exist, the findings tend to remain at the level of pedagogic implications only. Nelson's (2006) corpus-based materials, derived from a 1-million-word Business English Corpus, made up of both written and spoken genres, are the exception (see http://users.utu.fi/micnel/business_english_lexis_site.htm for corpus-based exercises targeting the genres of business contracts and minutes of meetings). DDL instructional materials have also been produced for teaching EU English, specifically the most frequent phrasal verbs and conjunctive cohesive devices (e.g. *with a view to, in the context of*) to professionals working in, or seeking a job at an EU institution (see Trebits 2009a, b). It should be noted that EU English, produced by native speakers or consisting of documents translated into English whose mother tongue is English, is not to be confused with Euro-English, a non-native variety of English (see Section 6.1.3).

On account of the lack of corpus applications in workplace domains, the discussion in the following sections is confined to the application of corpora for English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP) pedagogy, covering both 'hands-on' and 'pen-and-paper' DDL materials.

Corpora for EGAP

A key endeavour in the production of corpus-based materials to aid students with academic writing of a general nature is that by Thurstun and Candlin (1998a, b), as exemplified below. Moving from controlled to more open-ended

writing activities would seem to be inculcating in students the kind of 'corpus competence' outlined above.

Example 7.2 Corpus-based EGAP material

In this corpus-derived material the lexico-grammar is introduced according to its specific rhetorical function, e.g. referring to the literature, reporting the research of others. Within each broad function, each keyword (e.g. *claim, identify*) is then examined within the following chain of activities:

- LOOK at concordances for the key term and words surrounding it, thinking of meaning.
- FAMILIARIZE yourself with the patterns of language surrounding the key term by referring to the concordances as you complete the tasks.
- PRACTISE key terms without referring to the concordances.
- CREATE your own piece of writing using the terms studied to fulfil a particular function of academic writing.

(Thurstun and Candlin 1998b: 272)

It is quite surprising that since the publication of Thurstun and Candlin's 1998 textbook *Exploring Academic English: a Workbook for Student Essay Writing*, based on the 1-million-word MicroConcord Corpus of Academic Texts, there do not seem to have been any other similar initiatives, quite possibly due to the fact that producing such corpus-based writing activities is a time-consuming task. There exist a few generic materials for EGAP which are informed by corpus findings, but do not actually present any concordance lines for analysis. Feak et al. (2009) make use of transcripts from MICASE for interactions on campus, and Kelly et al. (2000) exploit video materials on lectures and seminar skills from the BASE corpus (see Section 4.4.2 for research on these two corpora).

In contrast to the dearth of corpus-based instructional writing material for EGAP, there are more reports in the literature on the use of corpus-driven, hands-on learning carried out institutionally (see L. Flowerdew in press 2012a for a review of these). However, in reality practitioners very often use a combination of both approaches. Like Thurstun and Candlin, Charles (2007: 296) also targets key rhetorical functions, in this case the combinatorial function of *defending your work against criticism*, a two-part pattern: 'anticipated criticism → defence and its realization using signals of apparent concession, contrast and justification'. Another feature of Charles's materials is that she approaches these functions by first using a top-down approach, providing students with a suite of worksheet activities to sensitise them to the extended discourse properties of this rhetorical function. She then supplements these

with a more bottom-up level approach by having students search the corpus to identify typical lexico-grammatical patterns realising these functions.

Corpora for ESAP

A variety of specialised corpora, consisting of lectures, engineering textbooks, legal essays and research articles, have been used for various types of pedagogic applications, which very often combine initial pen-and-paper awareness-raising activities with follow-up direct consultation of the corpus by students.

Jones and Schmitt (2010) devised discipline-specific vocabulary materials including both technical and colloquial items, derived from corpora of academic seminars on language and gender, international law and entrepreneurship (see Section 8.10). Mudraya's (2006) materials, based on a 2-million-word corpus of engineering textbooks, also targeted vocabulary, but of a sub-technical nature. Mudraya has noted that this type of vocabulary (i.e. those items such as *current, solution, tension* which have one sense in general English, but are used in a different sense in technical English) is problematic for students. She proposes a set of queries based around *solution* on the grounds that this word occurs, in its general sense, both as a high frequency word family and as a frequent sub-technical item. Students are presented with concordance output of carefully selected examples of *solution* and in one exercise are asked to identify, for example, the following: those adjectives used with *solution* (1) in the general sense and (2) in the technical (chemical) sense, and then asked to underline those adjectives that can be used with both senses of *solution* as a means to highlight collocational sensitivities.

Several pedagogic applications approach the corpus consultation from a genre-based perspective (Weber 2001; Bhatia et al. 2004; Noguchi 2004). For example, Bhatia et al. (2004) propose various move-specific concordancing activities for one genre of legal English, the problem–question genre written by students within academic settings. They note that deductive reasoning plays a major role in this highly specialised genre. One of their major foci, therefore, is to have students examine various types of non-lexical epistemic and pragmatic/discoursal hedges for the role they play in deductive reasoning.

Example 7.3 Step-by-step approach to exploiting the legal problem–question genre

Step 1: Awareness
If judgments and legal problem answers can be used with a concordancer, worksheets on hedging forms in legal problem questions and judgments can be used for students to:

– identify and classify hedges by function, form or grammar
– complete gapped concordance printouts on use of tenses, conditional clauses, etc.

– rank hedges by degree of certainty or doubt
– examine similar texts for variation in hedges used
– identify relationships between rule and analysis sections, etc.

Similar worksheets can be used with hard copies of the texts if a concordancer is not available.

Step 2: Contextualizing
Using samples from the rhetorical moves of answers to the same legal problem or of opposing appellate judgments, students could analyse and compare these for:

– typical lexico-grammatical structures and patterns
– variations in degree of certainty and doubt
– variation in rule statements and case citation methods
– appropriacy of conclusions based on deductive reasoning

Step 3: Application
– introduce students to various forms of lexical and non-lexical hedges
– work through high-frequency modals, conditionals, legal syllogism structures, case referencing methods, etc.

(Bhatia et al. 2004: 223)

Another advocate of a concordance- and genre-based approach to academic essay writing in the legal field, specifically formal legal essays written by undergraduates, is Weber (2001). First, Weber's students were inducted into the genre of legal essays by reading through whole essays taken from the University of London LLB Examinations written by native speakers, and identifying some of the prototypical rhetorical features, e.g. identifying and/or delimiting the legal principle involved in the case. They were then asked to identify any lexical expressions which seemed to correlate with the genre features. This was followed up by consulting the corpus of the legal essays to verify and pinpoint regularities in lexico-grammatical expressions. Similar to those tasks proposed by Bhatia et al. (2004), Weber also approaches the lexico-grammar from the perspective of a 'local grammar', which 'attempts to describe the resources for only one set of meanings in a language rather than for the language as a whole' (Hunston 2002: 90).

Concept 7.3 A local grammar for the genre of legal essays

Items such as *assume, consider, regard and issue*, in various constructions, were all found to act as signals in an opening-type move, delimiting the

(*continued*)

case under consideration before the principle involved in it was defined, as exemplified by the extract below.

received Brian's letter.	Assuming	the offer does remain open, Brian's Thursda
proceeding on the latter	assumption	In order to discuss the law related to
to discuss. I now have to	consider	whether B's message, left on the answerphone
third party. Bata v. Bata.	Considering	first the story about The BCDs, can the ba
With	Regard to	contracts *ex facie* illegal it is necessary to
second part of the story.	Regarding	the potential claim of Evangeline, it is subm
consented to it. The key	issue	here is what caused the injury; was it Geoffrey's
At	issue	here is whether Neil will have a cause of action

In an interesting departure from the normal type of ESP work, Weber's students were also exposed to corpora of different, non-legal genres in order to sensitise them to the highly specific use and patterning of certain lexical items, such as *held* and *submit*, in legal texts.

(Adapted from Weber 2001: 17)

However, as legal discourse is such an intricate discourse area corpus-based methodologies may not completely align themselves with legal writing tasks.

Concept 7.4 Corpus evidence vs institutional practices

Bhatia et al. (2004: 224) have underscored the complexity of legal discourse, pointing out that a number of academic and professional genres in law appear to be 'dynamically embedded in one another'. In view of this, they caution that one has to go beyond the immediate textual concordance lines and look at discursive and institutional concerns and constraints to fully interpret and by extension become a skilled writer of these highly specialised genres.

Hafner and Candlin (2007), in their report on the use of legal corpora by university students, also note a tension between professional discourse

practices which encourage students to focus on models, and the phraseological approach associated with corpus-driven learning. However, they see this as a tension to be exploited, arguing that 'continuing lifelong learners still need to be able to focus and reflect on the functional lexical phrases that constitute the essence of the texture of the documents they are composing' (p. 312).

A genre-based approach is also in evidence in the cycle of activities Bianchi and Pazzaglia (2007) adopted for helping Italian students to write psychology research articles in English. First, students were asked to subdivide their choice of a written article into moves and annotate it themselves using a functional and meta-communicative coding system devised by the authors. This was followed by data-driven guided writing tasks, which focused on the lexico-grammatical patterning of keywords related to the concept of research and verb tense usage in different moves. Another genre-inspired corpus initiative to aid postgraduate students in their writing of research articles in computer science is reported in Chang and Kuo (2011).

Swales (2002) contrasted the 'fragmented' world of corpus linguistics with its tendency to adopt a somewhat bottom-up, atomistic approach to text with the more 'integrated' world of ESP material design with its focus on more genre-based top-down analysis of macro-level features. However, Weber's tasks, in common with those of Bhatia et al., Bianchi and Pazzaglia, and those by Charles described in the previous section, now seem to be achieving a 'symbiosis' between these two approaches through their genre-based approach, as called for by Partington (1998) and absent from corpus-based pedagogy in the past (Swales 2002).

It can be seen that the pedagogic applications focus more on writing instruction than for speaking (see L. Flowerdew 2010). Also, direct uses of corpora have lagged behind indirect uses, i.e. for compiling dictionaries and grammars. Possible reasons for this lack of uptake are discussed below.

7.3 Potential impediments to DDL

It is evident from the previous discussion that most of the initiatives for integrating corpora into language learning have mainly remained at the institutional level and not filtered through to the language teaching community at large. Moreover, the vast majority of corpus work has been carried out with university students; the reports by Braun (2007) and Johns et al. (2008) on integrating corpus work into secondary education are rare exceptions. The following section mentions some obstacles which have contributed to the lack of uptake and recent developments that are overcoming these.

7.3.1 Corpora and tools

One possible reason for the lack of uptake of corpus-driven pedagogy by the teaching community is not the methodology so much as the medium. Some tools try to meet the needs of both researchers and teachers, which makes them overly complicated (see Breyer 2006a for a discussion on the needs of research and learner users as regards tools). This issue has also been flagged by Römer (2006), Krishnamurthy and Kosem (2007) and Granger and Meunier (2008: 251), who have indicated that one future challenge lies in 'creating ready-made and user-friendly interfaces to enable learners and teachers to access multiword units from a variety of genres and text types'. However, very recent endeavours are underway in this area and user-friendly tools, specifically to enhance academic writing skills, are described in Kaszubski (2011) and Milton (2004). Bloch (2009) describes a user-friendly interface for teaching the use of reporting verbs in academic writing. This interface presents users with only a limited number of hits for each query and a limited number of criteria for querying the database, namely integral/non-integral; indicative/informative; writer/author; attitude towards claim; strength of claim. Another key feature of all these tools is that they are accompanied by corpora compiled in-house to meet the needs of specific learners.

Another reason for the lack of uptake of corpus-driven pedagogy concerns the sometimes inappropriate nature of ready-made corpora for language learners. Naturally occurring corpus data would often be too difficult for low intermediate students to cope with and the tasks proposed by Bernardini (2002), while insightful for advanced learners, would be too challenging for lower level learners. As Osborne (2004: 252) comments 'Unless corpus examples are filtered in some way … many of the contexts are likely to be linguistically and culturally bewildering for the language learner'.

Example 7.4 Initiatives to identify difficulty level of corpora

Chujo et al. (2007) note that according to their readability index most of their English-Japanese corpora were rated at the advanced level, and they advocate that what is needed, therefore, are available e-texts at beginner level, of which there is a shortage. However, their study did identify several indices, e.g. word and sentence length, providing 'a rated collection of titles at varying levels of difficulty, and takes corpus usage one step closer to its ideal application' (p. 47).

Wible et al. (2002) describe a lexical filter which sorts examples according to a flexible threshold of lexical difficulty. A similar function is available in *SketchEngine*, which has an option to sort concordances 'best first' from a learner's point of view (Kilgariff, message posted on Corpus Linguistics discussion list 16 April 2008).

7.3.2 Strategy training for learners

Another impediment to the adoption of corpus-driven pedagogy by classroom practitioners may well be the fact that learners do not possess a pedagogic grounding in exploiting corpora, which may arise from the fact that the teachers themselves lack the necessary training (see Frankenberg-Garcia 2010). Interestingly, and somewhat surprisingly, there are very few accounts in the literature which touch on the question of learner training, but see Sripicharn (2010). One writing programme which has integrated strategy training into writing work is Kennedy and Miceli's (2002) exploitation of a corpus of contemporary written Italian to aid students with personal writing on everyday topics. As students were struggling with corpus queries Kennedy and Miceli gave directions for corpus investigations through a series of leading questions. This was followed up, after a few sessions, with the students encouraged to use the corpus on their own while revising their own work, the teacher acting as facilitator. These problems were the motivation for Kennedy and Miceli's (2010) 'pattern-hunting' and 'pattern-defining' guided tasks (see Example 7.6). O'Sullivan and Chambers (2006) report on how strategy training has been integrated into a writing course to help students improve their writing skills in French. See Flowerdew (2012b, in press) for another pedagogic application in which strategy training is built into corpus-driven activities to familiarise students with appropriate phraseologies for common functions in business letters (e.g. requesting, complaining).

Example 7.5 Strategy training for students

Lee and Swales (2006) provide a detailed overview with examples of exercises for inducting advanced-level students into the skills needed for exploitation of corpus tools and data. Students were introduced to the 'corpus way' of investigating language through, for example, using context to disambiguate near-synonyms and 'gaining sensitivity to norms and distributional patterns in language (semantic prosody; genre analysis)' (p. 62). One of these induction sessions is given in Figure 7.2.

Wk 6: Corpus, Usage Patterns and Subtle Nuances

Guessing/scrutinizing the meanings of words by studying concordances (e.g. *cabal*; *continually* v. *continuously*); Looking at similar lexical items (e.g. *for instance* v. *for example*; *effective* v. *efficient*; *expect* v. *anticipate*; *somewhat* v.*fairly*). Participant-generated examples of puzzling pairs, such as *totally* v. *in total*, *seek* v. *search*. …

Figure 7.2 Extract of induction exercises (from Lee and Swales 2006: 66)

Many practitioners seem to agree that induction-type tasks can present difficulties for students who are sometimes at a loss as to how to interpret the concordance lines. With this in mind, L. Flowerdew (2009) proposes consciousness-raising pedagogic mediation tasks in the form of hints to lead students toward interpretation of the corpus data. Meanwhile, Pérez Basanta and Rodriguez Martin (2007), after piloting their corpus of film transcripts for teaching rhetorical and discourse features of conversation such as requests, hesitation devices and backchannels, concluded that it was more beneficial to pre-teach these before corpus consultation to enable learners to more easily 'consciously work out the patterns by themselves and discover the social rules of the spoken situation' (p. 151). Gavioli (2002, 2005) also signals the potential dangers of induction-type tasks for students working with small, specialised corpora as they may overgeneralise their findings in a purely inductive-based approach and suggests that to counter this students compare results from different types of corpora to help them see the limits of their generalisations. All these reports thus support Johansson's (2009) preference for a 'guided-inductive' approach to corpus consultation (see Quote 7.6).

A third reason teachers may not have integrated corpus-based pedagogy into the curriculum is that, to date, there is very little empirical evidence to show the efficacy of corpus methodology, as discussed below.

7.3.3 Evaluation of corpus methodology

Although some very insightful studies have been conducted on learners' evaluation of corpora (Curado Fuentes 2002; Liu and Jiang 2009; Yoon 2008; Yoon and Hirvela 2004), much more empirical research needs to be carried out on the influence of corpus methodologies on learners' *performance*. Very few empirical research studies exist, which mainly focus on student writing and vocabulary knowledge (see Boulton 2007 for a comprehensive list of evaluation studies undertaken to date).

Cobb's (1997) research focusing on vocabulary shows improvement in students' lexical knowledge through implementation of hands-on concordancing techniques. Cobb (1999) also provides statistical data to show that corpus-based activities taking learners from a low intermediate stage of lexical growth up to functional reading can resolve the breadth–depth paradox of lexical acquisition. Students gain broad knowledge from their own self-compiled dictionaries based on corpora but at the same time are exposed to extensive reading for depth. Cobb's (1999) study relies on analysis of student logs to record actual corpus use, as does the study by Pérez-Paredes et al. (2011). This study is important for providing empirical data to show that skills and guidance are necessary for student consultation of a range of corpus-based resources.

Two empirical studies focusing on writing performance are those by Boulton (2009) and Cresswell (2007). Boulton (2009) has conducted tests on linking adverbials, and Cresswell (2007) on the use of connectors in experimental and control writing groups, which showed DDL in the context of the communicative teaching of writing skills to be moderately effective. Three studies which focus on students'

writing improvement in the revision stages of writing after being given feedback on errors are those by Gaskell and Cobb (2004), O'Sullivan and Chambers (2006) and Watson-Todd (2001). However, interestingly, an exploratory study by Jones and Haywood (2004) found that although students' awareness of formulaic sequences increased through corpus-based tasks, they did not do so well in transferring these phrases to their own writing. Thus the experimental results to date suggest that corpus consultation seems to be most effective for the revising process.

7.4 Under-represented corpora for pedagogy

It can be easily observed that both EGAP and ESAP corpora overwhelmingly consist of NS or 'expert' texts in English. Other types of corpora including learner corpora, corpora of other languages and multimodal corpora, still seem to be 'the poor relation' as far as pedagogic applications are concerned.

7.4.1 Corpora of other languages

Only a few studies exist which report on the use of corpora of other languages in foreign language teaching. For example O'Sullivan and Chambers (2006) describe how Master's students consulted a small semi-specialised corpus made up of two subcorpora, a 125,000-word corpus from *Le Monde* dealing with current issues relating to the French language and a 40,000-word corpus including text on the history and development of the French language, to improve their writing on a similar subject. Meanwhile, Kennedy and Miceli (2002) have built a 500,000-word corpus of contemporary written Italian, Contemporary Written Italian Corpus, CWIC, containing business letters, official e-mail messages and material from magazines, to aid students with personal writing on everyday topics. In a corpus study involving creative writing in Italian, Kennedy and Miceli (2010) describe their two-stage apprenticeship for inducting students into corpus consultation.

Example 7.6 Using a corpus of Italian for a creative writing project

Kennedy and Miceli's (2010) apprenticeship using CWIC consists of a 'pattern-hunting' followed by a 'pattern-defining' phase. For example, when writing about their sense of personal space for an autobiography, students were first prompted to come up with some key words for pattern-hunting, with many suggesting the common term *spazio*. This not only turned up ideas and expressions, e.g. *rubare spazio* (take space) but also triggered further searches on words encountered in the concordance lines, e.g. *percorso* (path). Other pattern-hunting techniques include browsing through whole texts on the basis of the title and text-type, and scrutinising frequency lists for common word combinations.

(*continued*)

The pattern-defining function was used when students did have a specific target pattern in mind to check. For example, one student wanted to establish if the pattern 'so' <adjective> 'that' could be rendered in Italian with *così* <adjective> *che* and if the subjunctive mood was required after *che*.

One large-scale initiative which addresses the paucity of corpora of other languages is the Aston Corpus Network (ACORN) project. One of the key aims of this project is to provide resources for the teaching and learning of languages (English, French, German and Spanish) across the University of Aston in the UK.

Although some valuable research studies have been conducted on French and Spanish corpora of academic and professional genres (cf. Lawson 2001; Parodi 2007; Tracy-Ventura et al. 2007), for the time being such findings tend to remain at the level of implications.

7.4.2 Multimodal corpora

Multimodal corpora based on SFL (see Section 4.5.1) are now making inroads into language teaching with a move away from monomodal to multimodal concordancing which takes other semiotics into account (cf. Ackerley and Coccetta 2007; Coccetta 2011).

Example 7.7 Multimodal functional-notional concordancing

Coccetta (2011) explains how students can make use of an online multimodal concordancer *MCA* (*Multimodal Corpus Authoring System*). This incorporates a search engine which allows students to find and isolate sequences in a corpus sharing the same characteristics by means of a functionally tagged corpus. For example, to see if the function of 'declining an offer' occurs in a subcorpus relating to *requests, invitations* and *offers* and to find the linguistic forms realising this function, students choose a parameter from a drop-down menu to retrieve the relevant concordance lines exemplified in Table 7.1.

Table 7.1 A set of results for the function 'declining an offer' retrieved with MCA

No thanks.
no thanks.
No thanks. I mean, that – that water's been there for ages.
No.
No thanks. I'm not – I'm not hungry.
Uh, no thanks.
I've already had one thanks.

Each concordance line has access to the film clip which provides non-linguistic information drawing on semiotic resources such as gesture, posture, gaze and facial expressions.

Another annotated multimodal corpus under development (but which is not rooted in Halliday's semiotic theory) is the SACODEYL suite of corpora (Widmann et al. 2011). These corpora consist of structured video interviews of 13–17-year-old secondary school pupils representing seven European languages: English, French, German, Italian, Lithuanian, Romanian and Spanish. The video material is aligned with the transcripts and the search engine allows for both topic-based queries and a search for relevant words and phrases for these topics. Thus, more pedagogic multimodal corpora of various kinds are now coming on-stream since the development of Braun's (2006) innovative ELISA corpus of interviews with video files constructed with clear pedagogic goals in mind.

7.4.3 Learner corpora

A great deal of valuable research has been carried out on oral and written learner corpora (see Sections 6.6.2 and 6.6.3). While most pedagogic applications of corpus findings involve native speaker or 'expert' corpora, several researchers-cum-practitioners (Gilquin et al. 2007; Granger 2009; Nesselhauf 2004), have also pointed out the value of incorporating learner corpora into language learning. Granger (ibid.) proposes two ways of doing this; for delayed or immediate pedagogic use.

Concept 7.5 Learner corpora for delayed or immediate pedagogic use

Corpora for delayed pedagogic use (DPU) are not used directly as teaching/learning materials by the learners who have produced the data. Rather, they are compiled by academics/researchers with the aim of providing a better description of one specific interlanguage and/or designing pedagogic activities for a similar population of learners with a similar profile. On the other hand, corpora for immediate pedagogic use (IPU) are collected locally by teachers with the learners being at the same time both producers and users of the corpus data.

(Granger: 2009: 20)

One experimental classroom project where IPU learner corpora have been integrated in the instruction cycle is reported in Mukherjee and Rohrbach (2006), who advocate individualising writing by having students build mini-corpora of

their own writing, and localising the database. A pedagogic initiative in which students compare a learner corpus of NNS MBA dissertation writing with a corpus of published journal articles from the field of Business Studies, both compiled by the teacher, is that by Hewings and Hewings (2002).

Example 7.8 Corpus-based materials derived from a learner corpus

Hewings and Hewings (2002) noted infelicities on the use of metadiscoursal anticipatory 'it' in student business writing. To draw students' attention to this they asked students to compare and discuss the differences of *it seems…* in concordance lines selected from the two corpora, as shown below in Table 7.2.

Table 7.2 Concordance task for *it seems…* in published articles and student dissertations

Published articles	**Student dissertations**
* It seems clear that as insider holding proportions increase; capitalization ratios decrease.	* It seems that different studies have shown different results.
* It seems likely that the eighties and nineties will be known as decades of large scale disaggregation.	* It seems that the practice of employing local staff by multinationals is increasing.
* It seems quite probable that consumers would not recognize such relatively small degrees of difference.	* It seems that some individual training courses are below their full capacity.

In spite of the acknowledged value in integrating learner corpora into language teaching, there is little evidence that learner corpora have had much impact on syllabus and materials design to date, an observation made by L. Flowerdew in 2000 and reiterated by Granger a decade later in 2009. It is expected that in the following decade the wealth of useful research data from the range of languages under the International Corpus of Learner English (ICLE) project (e.g. Ädel 2006; Hatzitheodorou and Mattheoudakis 2011) and findings from other learner corpora (e.g. Chuang and Nesi 2006; Lopez-Ferrero 2007; Lee and Chen 2009) will undergo transformation into pedagogic materials. While the impact of learner corpora has been felt in indirect applications, most notably to inform dictionary design (e.g. Paquot 2010), the same cannot be said in respect of direct applications.

Nevertheless, in spite of the potential advantages in integrating learner corpus data into pedagogy, Nesselhauf (2004) points out that care is needed

in presenting learner corpus data to students, as does Mukherjee (2009: 213): 'It is neither desirable or useful to establish a rigid dichotomy between good and correct usage in native data on the one hand and bad and incorrect usage in learner output on the other.'

Quote 7.5 Caveats in using learner corpora in data-driven learning

> ... since negative evidence naturally is only useful if the learner is aware that it is *negative* evidence, what has been called 'divergent learning' (Leech 1997: 11), i.e. browsing corpora to discover new facts about the language, is out of the question with learner corpora. Secondly, learners always have to be provided with positive evidence in addition to the evidence from the learner corpus. This evidence can come from either a comparable native speaker corpus or from a general native speaker corpus such as the BNC. Thirdly, since there is the danger that the learner will remember the negative but not the positive evidence or that the positive evidence will remain partly undigested (cf. Milton and Hyland 1999: 158), the exposure to negative evidence should be followed by exercises to consolidate the native speaker structure....
>
> (Nessselhauf 2004: 140)

A related consideration is for native language interference to be taken into account as this is also a source of error (see Granger 2004; O'Sullivan and Chambers 2006, and Section 6.6.4). This aspect is considered in a course for teaching technical writing using corpora compiled from the web (Foucou and Kübler 2000). Foucou and Kübler make reference to the L1 and point out that the use of the passive presents difficulties for French speakers as this construction is used less frequently in French than in English (e.g. '*On donne* ci-dessous des conventions pour ces options' would be translated as: 'Below, conventions *are given* for these options' (p. 67).

7.4.4 Corpora for L1 learners

The vast majority of corpus-based pedagogy is aimed at L2 learners of English, although there are a few reports of corpus work directed at L1 students, in particular young learners. A study by Sealey and Thompson (2004, 2007) and Thompson and Sealey (2007) investigated the use of corpus-based activities with primary school children at two schools in the UK, using 40 texts from the BNC classified as having been written for a child audience, i.e. imaginative fiction. Their CLLIP (Corpus-based Learning about Language in the Primary school) project exploited the corpus in several innovative ways, as illustrated below.

Example 7.9 Using colour-coding for word classes

Sealey and Thompson (2007: 215–16) used the CLAWS tag-set for the BNC to tag for different parts of speech, which they then colour-coded. The colour-coded concordance output stimulated discussion on which word classes children were sure about and which ones they were unsure about.

Extract 3: School B, Session 5
BB1 [489]: what are the brown ones?
BB1 [493]: the brown ones look like *cut*
GB1 [495]: like *kiss*
GB1 [497]: and *went, got, went*
GB2 [498]: I'm definitely sure they're verbs I think.

Function words, grammatical or non-lexical words emerged as a broad category of 'dull' words, which children had more difficulty with.

Sealey and Thompson's project was also tied in with the objectives of the National Curriculum, which specifies in Year 4 Term 2 that children are to be taught 'to spell words with the common endings: *ight* etc.' (National Literacy Strategy p. 40). To this end, Sealey and Thompson (2004: 84) produced a worksheet 'by instructing the concordancer to search for the string **ight*, and then selecting lines which would require the children to reflect on the relationship between this common letter string and the sounds associated with it'.

Reppen's (2001) longitudinal corpus study of the writing development of third to sixth grade elementary students also has important implications for pedagogy. Other corpora of L1 child language exist, the most well known being CHILDES which represents spoken language, but this has been used for computational and SLA-oriented analyses rather than pedagogic purposes (see Section 6.6.5).

7.4.5 ELF corpora

The role of English as a lingua franca in pedagogy, although still a controversial issue in some quarters (see Maley 2009), is gaining increasing acceptance. The most well-known ELF corpus is the Vienna-Oxford International Corpus of English (VOICE), which comprises speech events such as business meetings, casual conversations and service encounters among several others (see Section 6.1 for a discussion on research perspectives of this corpus). Besides the argument for accepting ELF in pedagogy on account of the fact that there is no reason for rigidly adhering to native-speaker norms when non-native varieties do not cause any miscommunication (Seidlhofer and Jenkins 2003), Mauranen (2004a), another proponent of ELF, seems to advocate its inclusion in pedagogy on account of affective factors.

Quote 7.6 Rationale for ELF

> … it is important for people to feel comfortable and appreciated when speaking a foreign language. Speakers should feel they can express their identities and be themselves in L2 context without being marginalized on account of features like foreign accents, lack of idiom, or culture-specific communicative styles as long as they can negotiate and manage communicative situations successfully and fluently.
>
> Holding up an NS model as the target for international users of English is counter-productive because it sets up a standard that by definition is unachievable. A more powerful solution is to capitalize on learners' strengths in acquiring and focusing on those aspects of the language that are relatively easy to learn (as core elements tend to be) and are most useful in communicating with other ELF speakers.
>
> (Mauranen 2003b: 517–18)

As far as pedagogic applications are concerned, van Rij-Heyligers (2007) has proposed building ELF corpora from the Web, as outlined below.

Concept 7.6 Arguments for using ELF corpora in EAP instruction

Van Rij-Heyligers (2007) is in favour of treating EAP as a lingua franca to convey the sense that academic genres are dynamic entities continually being shaped and negotiated by participants rather than as prescriptive and fixed artifacts. His EAP corpora were built using Web sources, which, he argues, are an ideal resource for compilation of ELF corpora as they reflect this changing nature of English. Moreover, van Rij-Heyligers (2007: 105) notes that corpora compiled from purely NS sources 'may contain the hidden message that the native speaker knows best, hence representing elements of linguistic imperialism'. Exploiting ELF corpora can thus be viewed as a resource to counteract the 'accommodationist' perspective on EAP writing in which students are often encouraged to adhere uncritically to conventionalised forms of expression.

Mur Dueñas (2009) and McKenny and Bennett (2011), who have carried out intercultural corpus-based analyses on research articles written in English and Spanish, and English and Portuguese respectively, also adopt a critical EAP pedagogy perspective and voice similar sentiments to those expressed by van Rij-Heyligers, arguing against a hegemonic approach.

However, the other side of the coin in the critical pedagogies debate is that by mastering the language of powerful genres, students can gain access to academic and professional discourse communities, thus gaining a sense of empowerment. This is Starfield's (2004) position, who uses concordancing with her students as 'a strategic engagement with technology' as a means of 'further exploration of issues of power and identity in academic writing' (p. 137). Starfield devised both worksheet and online concordancing activities as a consciousness-raising activity to foster awareness as to how writers position themselves with regard to the research of others, with a view to creating a niche for their own work and how they structure their own argument at the level of textual metadiscourse. She reported that students experienced a sense of empowerment when they realised they had readily available access to the language resources of authoritative English to expropriate these for their own means. A similar phenomenon has been noted by Lee and Swales (2006) in their students' reaction to the use of corpora as they realised they did not always need access to native speakers to check up on certain language issues.

Perhaps it is wise at this stage to advise caution, as Granger does, who also relates ELF to learner corpora and expresses doubts as to whether ELF features can be codified (see Concept 6.2). While the codification of a common core lingua franca may be arguable, native-speaker English of a non-standard variety, on the other hand, does display particular usage patterns. Anderson and Corbett (2010) make a case for using the SCOTS corpus (see Section 5.4.3) to raise ELT students' awareness of local speech varieties in English as a lingua franca.

Example 7.10 Pedagogic application of SCOTS corpus

Anderson and Corbett (2010) suggest focusing on what they term 'local words', characterising a particular native-speaker variety of English, to examine speakers' strategies for managing the pragmatics of conversation. They take the example of *wee*, which, in addition to its semantic value of 'small', also indicates a positive stance, often accompanied by laughter, e.g.:

A dug is a puppy it's a bonny *wee* beast. It takes ower yer hoose

Four other typical pragmatic markers of *wee* are noted:

1. to make a non-threatening offer, e.g. *would you like a wee seat*?
2. to downplay an imposition or bad news, e.g. *that's a wee shame*
3. to soften an insult, e.g. *a wee hairy*
4. with a delexicalised verb, to reduce pragmatic distance, e.g. *give a wee wave*

Anderson and Corbett suggest that after classroom discussion of how a local speech community accomplishes different pragmatic functions, teachers and learners could then reflect on alternative strategies for handling interpersonal features in conversation, either in standard English or in another ELF variety.

Of note is that the volume by Anderson and Corbett (2009: 162) is one of the few which builds a sociolinguistic dimension into the pedagogic tasks by asking questions such as 'Do younger speakers use more tag questions in speech than older speakers?'

7.4.6 Bilingual corpora

Bilingual, parallel corpora are more usually associated with translation research and training, and, as exemplified in previous sections of this chapter, different kinds of monolingual expert and learner corpora are the ones mostly used in language teaching. However, some practitioners have exploited parallel corpora for language teaching in monolingual classrooms where learners share similar L1-related difficulties (cf. Barlow 2000; Frankenberg-Garcia 2004, 2005a; Teubert 2004). Both Teubert and Barlow emphasise that parallel corpora are especially useful for examining phraseological queries, with Barlow noting that frequency counts provide 'a very good indication of the preferred structures in each language'. Frankenberg-Garcia (2005a) shows the value of using concordance output from a parallel corpus in preference to a bilingual dictionary as students can see the different contexts in which a word is used. She also sounds a note of caution, though, as decisions have to be made on the entry point for the search query. Foucou and Kübler (2003) also underscore the value of using bilingual corpora in ESP teaching, noting the difficulty French students have in learning verbs in the domain of computer science, and pinpointing deficiencies in existing textbooks and technical dictionaries.

Example 7.11 Using parallel corpora in language teaching

To take the case of 'false friends', Frankenberg-Garcia (2005a) notes that Portuguese learners of English frequently assume that words like *actually* and *actualmente* mean the same. In this case, she advises students use both L1 → L2 and L2 → L1 for the reason that 'looking up *actualmente* may help learners see that the equivalent in English can be rendered as *present, nowadays, these days, now* etc. …Looking up *actually* can help these same learners find out that it is a word whose equivalent is *de resto, na verdade*, or, most importantly, that it is often simply left out in Portuguese' (Frankenberg-Garcia 2004: 219).

One innovative application of a translation corpus is reported in Tim Johns et al. (2008) for teaching English through literature at a secondary school in Taiwan. A corpus of Arthur Ransome's *Swallows and Amazons* was used for devising corpus-based CALL programs. A Chinese translation of the text was also made available in order to give confidence to students with low English proficiency and to make tasks easier for such students who could then draw on their mother tongue for comprehension when needed. Another purpose for incorporating a translation corpus into the materials was to free students from the traditional one-for-one translation still sometimes found in traditional Taiwanese classrooms.

The remaining sections of this chapter discuss how corpora have been used in various content-based areas, namely in teaching translation, on teacher education programmes, and in teaching literature.

7.5 Corpora in teaching translation

Section 6.5 has discussed the theoretical framework underpinning corpus translation studies, the roles of different types of translation corpora (parallel, comparable, monolingual, disposable), and the use of corpora in research on universals of translation (simplification, explicitation, normalisation). This section looks at the application of corpora for teaching trainee translators in the academy.

Interestingly, Bernardini et al. (2003) take a wider view of translation training as a vehicle for language refinement for trainee translators. In fact, Fan and Xu (2002) in their evaluation of a bilingual corpus for the self-learning of legal English, found complicated syntactic structures to be an impediment to comprehension so this would certainly be another area for further investigation in translator training (see Case Study 8.9).

Quote 7.7 Corpora in language teaching and learning

> The uses of corpora in translation teaching contexts are not limited to translation classes proper, where corpus work can be relevant prior, during or after a translation task (Aston 2000). Corpora can also be used in other courses forming part of the curriculum of translation students, such as second language learning (Bernardini 2000) and terminology (Pearson 1998), and such work can complement translation activities in the narrow sense, developing capacities and competences that are far from marginal to translator education.
>
> (Bernardini et al. 2003: 4)

Besides the use of parallel corpora for language teaching purposes, the main aim of such corpora is in the teaching of translation, as noted by Beeby et al. (2009).

Concept 7.7 Using corpora in the teaching of translation

The intersection of corpora and translation teaching is seen from the following two perspectives:

- *Learning to use **corpora** to translate*, i.e. using corpora as tools and corpus linguistics as a method to find linguistic information useful in the translation process;
- *Learning to **translate** using corpora*, i.e. studying the process of translating using corpora, as in Castagnoli et al. (2011) who show how using a learner translator corpus in the classroom can lead to raising students' understanding of different translation strategies.

7.5.1 Learning to use corpora to translate

Kübler's (2011a) study addresses the first of the purposes above with reference to ESP tasks in a French-speaking setting. The corpora made available for students included a 1-million-word corpus of *Le Monde* newspaper (see O'Sullivan and Chambers's (2006) use of *Le Monde* for teaching French in Section 7.4.1), the English/French Europarl (European Parliament) parallel corpus (http://www.statmt.org/europarl) and a series of English/French and comparable corpora in earth science. The latter corpora can be described as disposable (see Varantola (2003)) as they are 'do-it-yourself' corpora put together by the analyst just for a particular task.

In the first task, first-year Master's students were required to work in groups to translate a research article in an imposed specific domain, i.e. earth science. The second task was a year-long individual project which required students to translate a text of about 5000 words in a specialised domain but of any genre. An overview of Task 1 is given below.

Example 7.12 Task-based approach to teaching translation

First-year master's students are assigned specialized research articles in two or three sub-domains of Earth Science, such as volcanoes, the birth and evolution of mountains, plate tectonics, hydrology, ice, climatology, and mud volcanoes. Each sub-domain and each article are assigned to a group of students. The articles are then divided into sections of about one thousand words, and each student is assigned one section. The aim of the project is to achieve a complete translation of the articles, with a consistent terminology for the sub-domain. The pedagogic objective is to lead students to discover the use of corpora in the process of translating a specialized text in a group translation context.

(*continued*)

The task is divided into a series of sub-tasks, some of which also relate to other courses. These are:

- Defining the genre;
- Collecting a corpus;
- Exploring the domain and understanding difficult or unknown concepts;
- Acquiring bilingual information on domain-specific terminology and phraseology, and on the phraseology of scientific argumentation;
- Conveying information appropriately in the target (native) language;
- Working together to agree on terminology and phraseology; and
- Revising the translation.

(Kübler 2011a: 69)

One interesting aspect of Kübler's study is that she illustrates how general corpora may often provide answers for specialised translations, when there is a lack of specialist articles in a particular domain, as illustrated below. Even when specialist articles are available on the internet the texts may be written by non-native speakers so the provenance and URL of such texts would have to be carefully scrutinised.

Example 7.13 Using general corpora for specialised translations

This example relates to a student's problem of how to translate *aggressively* in the following sentence from a computer science article:

> Attackers have discovered this vulnerability and are now *aggressively* index poisoning popular file-sharing systems.

Although the student's first inclination was to use *agressivement*, in the revision process he felt it to be incorrect so he checked it against the *Le Monde* newspaper corpus, which showed *agressivement* was used to modify adjectives rather than verbs. A search for *aggressively* in the Europarl corpus yielded the equivalent *de manière offensive*, which comes from the domain of war, typically collocating with *missile, armée, guerre, stratégie*, and was found with verbs describing a physical or verbal attack, an appropriate translation for the task at hand.

> Les attaquants ont découvert cette faille et *s'en prennent de manière offensive* aux indexs des réseaux P2P les plus populaires afin de les empoisonner.

(Adapted from Kübler 2011a: 75)

The above example shows that this task through learning by discovery is truly in the spirit of Tim Johns' data-driven learning where learners are encouraged to develop their own hunches and hypotheses about the data and devise their own strategies for searching and extracting data, accordingly (see Rodríguez 2006 for another course employing this data-driven methodology).

Pearson (2000) also gives useful suggestions for translator training by having students scrutinise the semantic relations in which the specialist term occurs.

Example 7.14 Searching for conceptual information in a corpus

When researching a term students are looking for a number of different categories of information. They will want to situate the term within its conceptual hierarchy: to do this, they will look for evidence of genus-species relations, part-whole relations, causal and other relations. They may want to identify the characteristics of the term, for example, its purpose, origin, and properties.

Genus–species relations
Frequently, when a term is introduced in a text, it is preceded or followed by its superordinate term or a general language equivalent:

> When a fairly severe virus known as The Ripper made its way into Ireland on one person's portable

Students can glean from the above that *The Ripper* is a virus. Other expressions which authors may use to indicate genus–species relations are *called, is a*.

Alternatively, a term may be introduced and explained in terms of its subordinate terms:

> The market for image setting systems (exposure units and film/chemistry systems) is expanding rapidly.

Specifying purpose, function, inputs, outputs, properties
When terms are explained, they are often described in terms of their function, properties or outputs alone. Phrases such as *used for, used to, has, have, is, are, involve, produce*, are commonly used to indicate that one or other of these types of information is being provided.

> 'Packaging' is defined as all products made of any material to be used for the containment, protection, handling, delivery and presentation of goods from raw materials to processed goods, from the producer to the user or the consumer.

(Adapted from Pearson 2000: 97–8)

7.5.2 Learning to translate using corpora

We now move to the second use of corpora in translator training, which is to study the process of translation with the aim of raising students' understanding of different translation strategies. In this respect, Frankenberg-Garcia's (2005b) study is of interest. While her research is not confined solely to corpora as a translation aid, nevertheless, the study merits attention for the integration of corpora with other sources for translator training and for understanding students' translation strategies to better inform translator training.

The students in this study were 16 final-year students taking a degree in translation at a Portuguese university, who were given a 200-word extract containing the first four paragraphs of a newspaper article about the effects the summer 2003 forest fires in Portugal would have on the country's ability to meet the Kyoto greenhouse gas emission targets. The students worked as individuals and had a wide range of online monolingual and bilingual resources to consult in the form of dictionaries and corpora, e.g. CobuildDirect Corpus Sampler and Compara, a bidirectional English–Portuguese corpus. The students were asked to make a note of everything they looked at by filling in a grid sheet (Table 7.3 as Concept 7.8).

Concept 7.8

Table 7.3 Recording translation strategies

What were you trying to find out?	What resource did you use?	What exactly did you look up?	Were the results helpful?	If so, what did you find out?	If not, why not?

(Frankenberg-Garcia 2005b: 338)

Analysis of the 146 look-ups recorded were grouped into the following five categories: finding an L2 equivalent; confirming a hunch; finding a suitable collocate; choosing the best alternative; checking spelling, with most queries centred on the first three types of look-up. A key finding of this study, though, was that 'the students generally preferred using resources mediated

by linguists, lexicographers and terminologists to resources requiring a greater amount of interpretation by the user…only 20% [of the look-ups] were performed using search engines' (Frankenberg-Garcia 2005b: 346–7). The bidirectional Compara corpus was only used three times. These results suggest that translators-in-training lack competence in integrating and utilising to their optimum the wealth of sources at their disposal and that strategy training is also needed just as it is for language learners in general (see Section 7.3.2).

The above discussion has shown that trainee translators have a lot to grapple with. Apart from familiarising themselves with the bounty of resources available, they also have to know how 'to do corpus linguistics' and understand the theoretical underpinning of the phraseological approach to language inherent in using corpora for translation purposes.

7.6 Corpora in teacher education

In Section 7.3.4 it was noted that the general lack of uptake of DDL by learners could be due to the fact that they lack strategy training in how to use corpora, and that this could be partly explained by the lack of training in how to use corpora by the teachers themselves. In fact, several practitioners (Frankenberg-Garcia 2010; McCarthy 2008; O'Keeffe and Farr 2003) have pointed out the gap that exists between the increasing development of corpus-driven, corpus-based and corpus-informed materials and pedagogy, and its inclusion in teacher education programmes. Questionnaire surveys aimed at secondary school teachers (cf. Heyvaert and Laffut 2008; Mukherjee 2004b) also indicate this lack of uptake.

However, in those teacher education programmes which do include a component on the use of corpora in pedagogy, some insightful observations have been made regarding the following three aspects, based on O'Keeffe and Farr (2003) and Farr (2010):

(i) teaching *about* corpora (technological awareness of what a corpus and concordancing are)
(ii) teaching *through* corpora (pedagogic awareness from analysing corpus samples)
(iii) teaching *with* corpora (linguistic awareness)

7.6.1 Teaching *about* corpora

Key notions to be covered here would include the different types of corpora available (spoken, written, multimodal, etc.), corpus design, size and representativeness. Teachers need to know how to choose among different types of corpora for

particular queries. Teachers would also be introduced to concordancing, a key analytical tool for corpus queries. Teachers need to know how to formulate different kinds of queries through specifying searches to the left and right of the node word and how to sort the concordance lines alphabetically. See Frankenberg-Garcia (2010) for a set of task-based consciousness-raising activities to familiarise teachers with basic corpus-searching strategies.

O'Keeffe and Farr (2003) also point out that teachers need training in how to 'read' concordance output, not only syntagmatically, i.e. from left to right, but also paradigmatically, from top to bottom.

Example 7.15 Training in 'reading' concordance output

… we find that learners need some training before they can make the most of concordance lines, including seeing collocational patterns. Reading a concordance line takes a little getting used to. The instinctive reaction is to try to read it in detail in the usual way, from left to right. We have found it best to skim it initially from top to bottom, looking only at the central patterns and working outward from them. For example, doing this with the concordance lines for *made* in Figure 7.3 reveals that it collocates frequently with a *case*, a *commitment*, a *decision* and so on.

lear. He believes it's important. He's	made	a commitment to get it done by the
rticularly sort of concerned with this,	made	a commitment at the beginning of t
of the lack of effort and they now have	made	a commitment. But they can answer
n intelligence activity in Bosnia. We	made	a condition of our train-and-equip
on who will listen with whom they have	made	a connection with in their freshmen
econd question. The statement has been	made	a couple of times that parents sho
tion. And in schools where they have	made	a decision not to use, they shouldn

Figure 7.3 Concordance lines of *made* sorted 1R and 2R

(Adapted from O'Keeffe and Farr 2003: 395)

However, as Frankenberg-Garcia (2010) points out, decoding corpus data may not be unproblematic for various reasons.

Quote 7.8 Problems with interpreting corpus results

Teachers who are used to dealing with language learning materials that have been carefully edited by experts cannot be assumed to know how to handle raw data samples containing language mistakes and idiosyncrasies,

> too many or not enough hits, and unforeseen findings that go against their expectations. We cannot presume that teachers will automatically be able to derive reasonable conclusions from all the information provided by the corpora without any help whatsoever. As corpus output is very different from the polished materials normally used in the classroom, teachers may need to be taught how to decode the results of their corpus searches.
>
> (Frankenberg-Garcia 2010)

Like Frankenberg-Garcia, Breyer (2006b) points out that teachers' IT competence, or lack thereof, and preference for more traditional resources are not to be taken lightly and that technological awareness is a key component of developing teachers' corpus competence. In other words, teachers need training in '*learning* to do corpus linguistics' (Kirk 2002: 156) such that they learn these basic processing skills from the perspectives of learner *and* teacher (Breyer ibid.).

7.6.2 Teaching *through* corpora

A necessary prerequisite for expert teaching is pedagogical content knowledge consisting of content knowledge (i.e. linguistic knowledge in the case of EFL teachers), pedagogic knowledge and content-specific teaching knowledge (Shulman 1987, cited in O'Keeffe and Farr 2003).

Concept 7.9 Pedagogic and content-specific teaching knowledge

> Pedagogic knowledge covers strategies for motivating students and managing the classroom environment, such as the use of questioning techniques, nominating students, giving instructions and grouping students. Content-specific teaching knowledge, on the other hand, concerns the application of pedagogic knowledge to specific sociocultural and organisational settings.

Pedagogic and content-specific teaching knowledge have both been addressed in corpus-based modules on teacher education programmes. O'Keeffe and Farr (2003) outline a series of tasks for raising students' awareness of pedagogic knowledge through analysis of corpus classroom data. It is of interest to note that they combine this aspect with raising teachers' technological awareness and also content knowledge of discourse analysis by building hints

on searching into the instructions, and by asking teachers to analyse the concordance output in terms of Sinclair and Coulthard's (1975) model of classroom discourse. They also point out that the corpus data chosen was from both expert and non-expert teachers (instead of experienced versus inexperienced teachers) to avoid equating inexperience with lack of expertise and vice versa. Moreover, O'Keeffe and Farr's sample material for raising awareness of pedagogic knowledge in Example 7.16 demonstrates considerations that need to be taken into account for data entry of spoken discourse; the person responsible would need to recognise which of the spoken utterances signal a question and enter '?' accordingly.

Example 7.16 Acquiring pedagogic knowledge through classroom corpus analysis

... trainees start by looking at the questioning patterns in our classroom corpus. They investigate the correlation between a question type and its productivity (they quickly notice, e.g. how much more productive referential questions are than *yes/no* questions). They are then asked in Task (c) to look more broadly at the placement of *questions* + *response* + *follow-up* for each question type within Sinclair and Coulthard's (1975) initiation–response–feedback model (see Figure 7.4).

a) Run concordances of questions used in the classroom corpus to determine their frequencies. (wh-questions can be extracted by searching each of the *wh*-questions individually, and *yes/no* and intonation questions can be found by searching '?')
b) Analyse and compare the productivity of each question type by running an analysis of student responses in terms of length and quality (use up to 10 examples of each question type).
c) How does each type fit in the typical initiation-response-follow-up (IRF; Sinclair and Coulthard 1975) classroom exchange structure? Use the KWIC* facility to help with your analysis.

* Keyword in context. Instead of viewing only short concordance lines, students can view an extended context for each occurrence of the search term.

Figure 7.4 Sample material based on the Limerick Corpus of Irish English for raising awareness of pedagogic knowledge

(Adapted from O'Keeffe and Farr ibid.: 398–9)

Another initiative which aims to raise trainee teachers' pedagogic awareness through investigation of the function of discourse markers used in classrooms

teaching French and Spanish is that by Amador Moreno et al. (2006). Their discussion with the student teachers centred on the multifunctionality of some of the discourse markers and their different linguistic and pragmatic functions (Table 7.4).

Example 7.17 Pedagogic functions of discourse markers in French and Spanish

Table 7.4 Functions of discourse markers

Function	Discourse markers
1. Introduction of new topic, activity, or question	French: *alors*, *bon*, *OK*, *bien*, *donc*, *Allez* Spanish: *vamos a ver*, *bueno*, *vale*
2. To call the pupils' or teacher's attention	French: *d'accord*, *bon* Spanish: *vamos a ver*, *a ver*, *mira*, *oye*
3. To recap what has been said or offer clarification	French: *alors*, *donc*, *d'accord* Spanish: *o sea*, *vamos a ver*, *entonces*
4. To motivate or encourage the pupils	French: *allez*, *alors* Spanish: *anda*, *venga*, *va*
5. To correct oneself or rephrase what has been said	French: *c'est-à-dire*, *enfin*, *'fin* Spanish: *o sea*, *bueno*

(Amador Moreno et al. 2006: 92)

Amador Moreno et al.'s data are made up of three subcorpora: non-native speaker teachers of French/Spanish teaching the language to non-native speakers in Ireland; native speakers of French/Spanish teaching the language to non-native speaker pupils in Ireland, and native speakers of French/Spanish teaching the language to native speaker pupils in a country where the language is spoken. Their rationale for including the third sub-corpus was, as well as focusing on 'teacher talk', to also include a focus on 'student talk': 'The third category is included primarily as it provides the student teachers with examples of the language use of native speaker pupils, which may be of use to them when offering guidance to their pupils on using the target language in the classroom' (p. 86). Their study did not set out to explicitly compare the discourse markers used by native and non-native speaking teachers, but it was noted that the native speaker language teachers used a relatively small number of discourse markers with a limited number of functions, '…thus making mastery of their use an achievable goal for the nonnative speaker teacher' (p. 99).

Cameron's (2003) research on metaphor in teacher talk could well be of value in informing teacher-training programmes at the primary school level.

Example 7.18 Metaphor in educational discourse

Cameron and Deignan (2003: 153–4) report on the use of 'tuning devices' (i.e. *just, like, sort of*) around metaphors found in a small intensively studied corpus recorded in a primary (elementary) school in the UK. Tuning devices accompanying metaphors show these to have the following two main functions:

Directing listeners to a particular interpretation:
To prevent a metaphor from being understood literally: e.g.

… ***a sort of*** *'nickname' (.)* ***a sort of*** *'corruption.'*

For the reverse microfunction, i.e. to prevent a metaphorical interpretation of a statement, indicating a literal meaning: e.g.

… just imagine rock getting so hot that it ***actually*** melts.

Adjusting the strength of the metaphor:
To tone down the potential strength of a metaphor or mitigate its implications: e.g.

… it just looks like **a kind of** '*shuffle*.' (boys' dancing)

Cameron and Deignan interpret the fine tuning devices around metaphors as used for management or pedagogic purposes, and state that metaphors used on their own may sound rather blunt or strong, and may be open to misinterpretation in the context of the classroom.

Yet another type of corpus offers valuable insights for inclusion in teacher education programmes, corpora of recorded training sessions, i.e. pre- and post-observation conversations between mentors and trainees. Such corpora would serve to raise trainees' awareness of content-specific teaching knowledge through reflective practice. For example, Farr (2003, 2008) examined interactional language between mentors and teacher trainees in a 60,000-word corpus of Post-Observation-Teacher-Training-Interactions (POTTI), finding that hedging expressions of possibility and conditionality were used to couch, what in effect are directives, as suggestions, e.g. *If I were to do this exercise I would approach it from an elicitation point of view* (O'Keeffe et al. 2007: 129), a device also used in a health care context for mitigating directives (see Section 6.2.2). In another study Vásquez and Reppen (2007) used the corpus data from their teacher feedback sessions to change the conversational approach of the mentors, thereby producing more extended and better quality feedback.

Walsh (2006) has compiled a reflective feedback corpus of EFL tutor comments to foster a fine-grained critical self-evaluation of tutors' own teacher talk, which draws on aspects of sociolinguistic theories such as conversation analysis (see Section 3.4.1). He first drew up a framework for analysing classroom interaction consisting of four pedagogic goals, managerial, materials,

skills and systems, and classroom context, against which are mapped a series of interactional features, e.g. scaffolding, direct repair, extended wait time etc. (cf. Walsh ibid.: 66–7). The teacher self-evaluations are phrased in terms of these modes and interactions, as exemplified below.

Example 7.19 Extract from stimulated recall corpus

Classroom interaction	Teacher's commentary
T ok now which of those foods do you prefer? (4)? L oysters and olive oil= T =and olive oil yeah which do you Prefer Jason? L (5) T what? which food do you prefer from all of this? (3) Kevin which food do you prefer? L oh yes. Deep-fried.	*That's classroom context mode now. I know that they know the words that they're looking for but they can't quite get it out so I give them the time, a few seconds, enough time to let them get the word out rather than pass them over and not give them the opportunity to say the word that I think they know. This is a useful strategy because it builds up their confidence if you give them that time.*

(Walsh ibid.: 133)

Walsh's reflective feedback corpus therefore seems to be targeting both pedagogic knowledge, through the teachers' identification of interactional features, and content-specific teaching knowledge through teachers' reflection and commentary on how these interactional features relate to their own sociocultural classroom environment. Another corpus focusing on reflective practices which shed light on the teacher education process is that compiled by J. Flowerdew (2002a), of learner diaries of students on a BA TESOL course.

Thus, there exist quite a number of individual studies using consciousness-raising activities to foster pedagogic awareness through corpus analysis. Two of these studies report on the teachers' evaluation on the use of corpus-based instruction in teacher education programmes (Amador Moreno et al. 2006; Farr 2008), which is generally favourable. However, the trainees in Amador Moreno et al.'s study expressed a strong preference for video recordings on account of the fact that non-verbal gestures are often used to reinforce discourse markers. Multimodal corpora will no doubt make more of an appearance in future work, in line with advances in discourse analyses (see Section 4.5) and pedagogic applications (see Section 7.4.2), utilising multimodal corpora of various kinds. Further research in the form of corpus-informed longitudinal studies evaluating the entire teacher training process which could also incorporate ethnographic data from interviews and diaries would serve to provide empirical data on trainee teachers' performance and to what extent they benefited from corpus-based insights and activities.

Corpora have also been used in teacher education programmes for raising teachers' own linguistic awareness.

7.6.3 Teaching *with* corpora

Teaching with corpora to raise teachers' linguistic awareness was first introduced in teacher education programmes in the mid-1990s, together with training in using corpora (cf. Coniam 1997a; Hunston 1995; Renouf 1997). These studies emphasise the benefits of corpus-based enquiries to focus on phraseological patterns or semantic information which may not be found in grammars and dictionaries. The philosophy underpinning such enquiries is: 'The learner should be aware of grammar as a method rather than as a set of facts; that grammar is about knowing how to observe and how to interpret observations, rather than knowing what other people have observed' (Hunston ibid.: 16).

Example 7.20 Interrogating a corpus

There are many questions that cannot really be answered without recourse to a corpus. For example, in a course focusing primarily on lexis, one might pose such questions as:

1 What are the core or primary meanings of the words *keep* and *see*?
2 What would you regard as the main facts about the meaning and use of the word *listen*?
3 How much can you say about the meaning and use of the words *affirm* and *confirm*?
4 Give examples of how the following words are used in text: *axiomatic, synergy, symbiosis, serendipity*.

(Adapted from Renouf 1997: 259)

(Trainee) teachers are advanced language users so it is not surprising to find that corpus-based work focuses on honing their reflexive and critical skills in various ways. For example, Seidlhofer (2000, 2002: 226) got her students to compile their own corpora of summary writing and then used student-generated questions for more sophisticated analysis of lexis and variation, e.g. *Did we use synonyms for words in the original articles*? *Which words are probably taken from the text* (*as they are not common vocabulary*)? *Which other vocabulary do advanced learners tend to use*? The students in Farr and O'Keeffe's (2002) programme examined grammatical choices from a sociolinguistic perspective by comparing the use of *would* across three corpora: British and American English presented in the LGSWE (Biber et al. 1999), with Irish English in the Limerick Corpus of Irish English, L-CIE,

and New Zealand English in the Wellington Corpus of Spoken New Zealand English. Sensitising student teachers to different varieties of English also reignites the issue of English as a lingua franca. In this respect, Sifakis (2007) argues for incorporating material from published spoken ELF corpora into teacher training materials on the grounds that ELF is now emerging as a codified variety in its own right (but see Section 7.4.1 for contrasting viewpoints).

Meanwhile, Osborne (2000) devised tasks which encouraged students to question the rules found in pedagogic grammars.

Example 7.21 Comparison of corpus findings with pedagogic grammars

In order to engender a sense of critical awareness in his students and confront them with findings that might run counter to the rules commonly found in pedagogic grammars, Osborne (2000: 170) provided his students with corpus examples of modal and periphrastic futures. These were accompanied by a series of prompt questions about 'frequency, context, possible meanings, differences between problematic areas of usage etc.' Students were asked to consider the following:

- *will* is more frequently used than going to.
- *going to* is used to talk about an immediate future.
- *going to* is not used in future conditional sentences.
- *going to* is used to talk about things that have already been decided.
- *will* is used to talk about decisions and promises.

It has also been proposed that corpora are useful for non-native teachers to consult when they either want to corroborate or disconfirm native speaker judgments, especially when these are in conflict (Mair 2002), or they want more information on a specific language structure raised in the classroom. In this respect, Tsui (2004, 2005) describes how practising teachers can access a website, Telenex, to consult corpus resources for queries their students bring up in class, such as the difference between *big* and *large*.

7.6.4 Corpora in EAP/ESP teacher education

Specialised corpora have also been used on teacher education programmes for informing the teaching of EAP and ESP. For example, Coxhead and Byrd (2007) describe how a variety of corpus resources can be used to prepare writing teachers to teach the vocabulary and grammar of academic prose. ESP subcorpora form part of a corpus for use in Chinese language education (Szakos 2000).

A framework for mediating the findings of corpus linguistics and genre analysis on a TESP module is outlined in Hüttner et al. (2009), who note the general neglect of ESP training on teacher education courses.

Example 7.22 Teaching-oriented corpus-based genre analysis for ESP

Hüttner et al. (ibid.: 106) propose the following framework (Table 7.5) to familiarise student teachers with the potential of specialised corpora as a resource for investigating specific genres and as a tool in materials development.

Table 7.5 The investigative procedure of teaching-oriented corpus-based genre analysis

(A)	Selection of genre and description of the teaching situation envisaged (incl. the imagined group of learners)
(B)	Description of the genre (communicative purpose/s + potential discourse community)
(C)	Collection of exemplary texts and compilation of the mini-corpus
(D)	Description of the 'moves' on the basis of the texts included in the mini-corpus
(E)	Lexico-grammatical analysis: comparison of mini-corpus with reference corpus (BNC) with WordSmith tools
(F)	Analysis of textualisations: connecting investigative steps (D) and (E)
(G)	Interpretation of the results with reference to the teaching and learning situation, developing teaching materials

7.6.5 Concluding remarks and future directions

While the use of corpora in teacher education programs is becoming increasingly common (see the special issue on Teacher Education in the *International Journal of Corpus Linguistics* (Farr and O'Keeffe 2011), one point to note is that the studies discussed above focus on corpora of *face-to-face in-class* interactions.

With the increasing globalisation of education and hence the proliferation of more *online distance* education programmes, this type of interaction could be analysed to inform teacher feedback. For example, while not explicitly relating their findings to teacher education, Hewings et al. (2009) analyse a corpus of e-conferencing from an academic writing programme. Computer-mediated communication is a burgeoning yet under-researched field (see Section 4.6), which is ripe for analysis also by teacher educators. Another area that could be fruitfully addressed on teacher education programmes is that of intercultural communication (see Section 6.2.1). In this respect, Reinhardt (2010) has analysed a corpus of ITA (International Teaching Assistant) interactions from an intercultural perspective.

7.7 Corpora in teaching literary analysis

In spite of the emergence of corpus stylistics as a well-defined field in research (see Section 6.4), there are just a few endeavours reported on teaching literature using various corpus stylistic methods.

7.7.1 Initiatives in using corpora

Having students focus on collocations figures prominently in pedagogic applications, with Bednarek (2007) touching on feminist issues and Kettemann and Marko (2004) on CDA. Bednarek (ibid.), based on frequent three-word clusters with negation in Charlotte Brontë's *Jane Eyre*, got her students to check the collocations of the negations. She concludes that the concordance evidence with negation can be interpreted as construing the female protagonist as 'rather independent, strong-willed, and courageous ... She refutes stereotypes against women and does not conform to normal beauty standards that we expect of women' (p. 8).

Section 6.4.3 has shown the central role that corpora have played in analyses of speech representation in literary research, an aspect also addressed in pedagogic applications. Using a corpus of Shakespeare plays tagged for parts of speech, Kettemann and Marko's students carried out searches on performative verbs, which were found to feature quite prominently. A search on 'say' used as a performative verb revealed that very often it collocated with imperatives (e.g. *Call hither, I say, bid come before us Angelo*). As the collocations suggest that 'say' has the function of adding more emphasis to an order, students could conclude that 'those characters having their *say* are those who are in power' (p. 181).

One innovative endeavour in this field making use of 'new technologies' is that by Kehoe and Gee (2009), who have devised a wiki tool which builds upon the WebCorp Linguist's Search English (Renouf 2003; Renouf et al. 2007).

Concept 7.10 Wiki tool for corpus-assisted, collaborative study of literary texts

> The wiki interface allows teachers and students to attach analytical and interpretative comments to individual words and phrases, and thus functions in a similar way to the tracking functions in MS word. Moreover, these comments can generate intra- and inter-textual links. Concordancing and statistical analyses are also incorporated into the software.
>
> (Kehoe and Gee 2009)

7.7.2 Arguments against a corpus-based approach

However, Kettemann and Marko (2004) have raised some strong arguments against a corpus-based approach to teaching stylistics.

Quote 7.9 Arguments against a corpus-based approach to stylistics

Destroying the integrity and wholeness of texts: Using concordances means fragmentizing texts (treating it as *text* – uncountable noun – rather than as *a text* – countable noun), thus destroying the integrity and wholeness of the text and thus also suspending interpretation triggered in a linear reading of texts.

Promoting uncritical and superficial reading of texts: Using concordances promotes an overemphasis on surface form at the cost of deeper meanings, which precludes the possibility of resistant and critical readings of texts.

Blurring literary issues: Corpus analysis of literary texts may help to study literary texts linguistically, but won't be able to answer the genuine questions of literary studies. In other words, corpus analysis promotes an approach indifferent to the literariness of literary texts.

(Ketteman and Marko 2004: 190)

Of course, the above criticisms would stand if 'corpus analysis is promoted as the *only* analytical instrument in the study of literary text' (Kettemann and Marko ibid.: 190). For the teaching of literary studies, it seems that the wiki tool developed by Kehoe and Gee (ibid.) would serve as a means of both promoting the linear reading of whole texts and using corpus-based methodologies to hone in on features of interest.

Further reading

Bennett, G. (2010) *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Ann Arbor, Mich.: University of Michigan Press. This practical guide serves as a user-friendly introduction for teachers.

Campoy-Cubillo, M., Bellés-Fortuño, B. and Gea-Valor, M. (eds) (2010) *Corpus-Based Approaches to English Language Teaching*. London: Continuum. This volume contains papers which report on the use of corpora in ESP courses.

Gabrielatos, C. (2005) Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ*, 8 (4): 1–37. This article presents a very useful overview of the application of corpora in language teaching.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. London: Routledge. This comprehensive volume approaches the field from an applied linguistic perspective.

Reppen, R. (2010) *Using Corpora in the Language Classroom*. New York: Cambridge University Press. This introductory volume presents a useful overview of the area.

8
Research Cases

This chapter will:

- Present and evaluate ten examples of corpus-based research projects which adopt different analytical perspectives to the data
- Use these cases to illustrate how various corpus-based methodologies and tools have been applied to the analysis of a wide variety of genres, contexts, sites and domains
- Examine some exemplary research practices of contemporary corpus-based research
- Suggest how researchers might develop and adapt the methods and results of these cases for their own research projects.

The cases discussed in this chapter either focus on small-scale projects using specialised corpora, or the investigation of specific phenomena from large-scale corpora. Given their design and scope, they offer valuable suggestions to readers (both students and teachers) working as individuals, pairs or small groups. Each case begins by providing a context for the research and is followed by a summary of its aims, corpus, methodology, results and analysis, followed by a commentary on its design and the contribution the study makes to the field of corpus linguistics.

I conclude each case by mapping out areas for further exploration, some of which are suggested by the researchers themselves. Readers are invited to consider how they could remodel these cases for their own situations either through replication studies or through modification in some way.

8.1 Comparison of oral learner and native-speaker corpora from a phraseological perspective

Summary 8.1

De Cock, S. (2011) Preferred patterns of use of positive and negative evaluative adjectives in native and learner speech: an ELT perspective. In A. Frankenberg-Garcia, L. Flowerdew and G. Aston (eds) *New Trends in Corpora and Language Learning*, pp. 198–212. London: Continuum.

This study investigates evaluative adjectives (e.g. *good, great, nice, bad, awful, terrible*) in two corpora: the Louvain Corpus of Native English Conversation (LOCNEC) and the French, German and Chinese components of the Louvain International Database of Spoken English Interlanguage (LINDSEI). The analysis focuses on the preferred syntactic and collocational patterns in which these adjectives tend to be used. Implications for ELT pedagogy are presented and discussed.

To date, most research comparing native and learner corpora has been carried out on written academic corpora of a general nature (see Section 6.6.2). It is only recently that more attention has been paid to differences in speech. This article reports on an investigation of attitudinal stance, which conveys speakers' attitudes, likes or dislikes, or evaluations of events or personal experiences. The focus of the study is on the adjectives that native speakers and advanced EFL learners use recurrently to express both positive and negative evaluation, and draws on data from the LOCNEC and LINDSEI corpora respectively.

8.1.1 Aims

The main aims of this research are twofold:

1. To paint an overall picture of the use of frequently recurring positive and negative evaluative adjectives in native speaker speech and in the spoken productions of advanced EFL learners from Chinese, French and German mother-tongue backgrounds;
2. To analyse from a contrastive perspective the preferred syntactic and collocational patterns in which the adjectives are used in the three varieties of English.

8.1.2 Corpora and methodology

Learners' use of evaluative adjectives was analysed using three (Chinese, French and German) out of the 11 varieties of learner English from the LINDSEI project (see Granger 1998a). Fifty learners for each variety contributed data to LINDSEI (see http://www.uclouvain.be/en-cecl-lindsei.html for further details). These

learners are labelled as advanced on the basis of an external criterion, i.e. they are third and fourth year students at an English university. LOCNEC is made up of informal interviews with 50 British university students. The corpora totalled between 70,000 and 110,000 words of interviewee speech respectively.

The corpora parallel each other in several ways. The informal interviews are of similar length and follow the same set pattern. An informal and open discussion, mainly centred around topics such as university life, hobbies, foreign travel, or plans for the future, forms the major part of the interview. De Cock points out that a short picture-based story-telling activity was included at the end of the interview to allow for *targeted comparisons* of lexis between various learner varieties or between learner and native-speaker varieties. She also notes that the types of topics included in the informal interviews in which learners are encouraged to express personal attitudes and feelings, lend themselves well to a study of evaluative markers.

The positive and negative evaluative adjectives were first identified on the basis of frequency lists of word forms from the corpora using *WordSmith Tools 4.0*. The frequency threshold was set to a minimum of 10 occurrences per 70,000 words. As for meaning, the adjectives selected all had to be considered as prototypically evaluative and fit into Biber et al.'s (1999) evaluative subcategory of descriptors, i.e. adjectives denoting judgements, affect or emphasis. Adjectives prototypically used to indicate size, e.g. *big* and *little* were excluded, as were those not used for evaluation, e.g. *I didn't think of that. erm. yeah for a while . but not for* ***good*** *because erm. I love my . country*.

8.1.3 Results and analysis

The frequency counts indicated that the positive evaluative adjectives under scrutiny tended to occur with much higher frequencies than the negative evaluative adjectives, as shown by the five most frequent adjectives in Tables 8.1 and 8.2. Moreover, it was found that some of the most frequently recurring evaluative adjectives could be traced to one specific part of the interview, namely the short, picture-based, storytelling activity.

De Cock explains the positive bias in both LINDSEI and LOCNEC in the following way. The interviewees could have been influenced by the first two topics discussed at the beginning of the interview, which could be regarded as tending towards the positive. Alternatively, the positive bias in the corpora could be linked either with the interviewees' desire to make a positive impression on the interviewer to establish rapport, or the interviewers may have been attempting to use more positive language to put the students at ease. It is also of interest that the adjective *ugly* occurs far more frequently in LINDSEI_CH than in the other varieties under study. Xiao Chen (cited in De Cock 2011) proposes the following possible explanations: (1) one of the Chinese equivalents of *ugly* is rather soft and not as strong as the English

Results 8.1

Table 8.1 Frequently recurring positive evaluative adjectives (relative frequencies per 70,000 words) (adapted from De Cock 2011: 202)

	LINDSEI_CH		LINDSEI_FR		LINDSEI_GER		LOCNEC (NS)	
1	good	231	good	85	nice	152	good	219
2	beautiful	228	beautiful	80	good	146	nice	122
3	Satisfied	63	interesting	71	beautiful	57	interesting	39
4	happy	53	nice	61	interesting	48	interested	36
5	important	50	great	39	impressive	40	happy	30

Table 8.2 Frequently recurring negative evaluative adjectives (relative frequencies per 70,000 words) (adapted from De Cock 2011: 203)

	LINDSEI_CH		LINDSEI_FR		LINDSEI_GER		LOCNEC (NS)	
1	ugly	50	difficult	46	difficult	48	Difficult	40
2	angry	37	awful	31	hard	39	bad	38
3	difficult	26	bad	19	bad	36	weird	12
4	bad	24	ugly	14	angry	15	horrible	11
5	wrong	19	disappointed	11	ugly	15	hard	10

adjective; (2) *ugly* appears to be commonly taught as the antonym of *beautiful* without mentioning its 'taboo' flavour, while other milder antonyms of *beautiful* are rather poorly represented in teaching materials; (3) when performing the picture-story task, the Chinese learners may have tried to be imaginative and used *ugly* for dramatic effect.

De Cock also reports that a closer scrutiny of the context reveals that the binary division into positive and negative obscures a more complex situation: while the adjectives *nice*, *great, interesting, awful* and *hard* tend to be used mainly in assertive contexts, others are used non-assertively with the adjective *bad* often found in the pattern *BE not (too/that/so/as) bad* in LOCNEC and LINDSEI_GER.

Differences in syntactic and collocational patterning of these adjectives in LOCNEC and LINDSEI were noted. One preferred syntactic environment in the native-speaker corpus was the evaluative relative clause introduced by *which*. Many of these were sentential clauses commenting on the whole previous sentence, series of clauses or utterance, e.g. 'first of all the way that the informants speak English *which is **great** for me* cos I speak no Dutch at all'. However, evaluative (sentential) relative clauses were a far less preferred syntactic environment in the three learner varieties. The preferred positioning for evaluative adjectives was inside an NP before the head noun, e.g. 'it was a *very **good** experience for me . yes because they . made a lot of critics*' (…) (LINDSEI_FR).

8.1.4 Commentary

This is an exemplary corpus-based study of three different learner varieties of spoken English compared with a corpus of native-speaker speech. The corpora have very clear design criteria which allowed for each to be paralleled with the others to enable targeted comparisons. One aspect of corpus design concerns whether external or internal criteria are used for classification purposes. An external criterion of year of study was used for labelling the learners as advanced. However, internal criteria may also have a role to play. De Cock points out that, in reality, raters grading the interviews according to descriptors using the Common European Framework for Languages may grade the Chinese learners as higher intermediate and the other learners as advanced, as was the case for the written component. Corpora are usually constructed with a priori questions in mind and these corpora of interviews are well suited to a study of evaluative lexis. This study also shows that frequency lists may need some pre-processing before subsequent analysis, as the lists of adjectives had to be trawled through to discard any adjectives not used for evaluation. De Cock has set the cut-off point to at least 10 occurrences per 70,000 words to capture the most prototypical evaluative adjectives. One thorny issue regarding frequency counts is how to determine the cut-off point for subsequent follow-up analysis; these necessarily usually involve some degree of subjectivity as in this case and the majority of corpus-based analyses.

Interpretation of corpus findings may not be an easy matter, with several explanations possible. Here, De Cock is suitably tentative in providing explanations for her corpus findings, which have been related to the task, topic and semantic equivalence across languages in the case of the high use of the adjective *ugly* in LINDSEI_CH. Several other applied corpus linguists (see Section 7.2.2) have noted the influence of instructional materials on learner writing and De Cock makes a similar observation for speaking with regard to the recurrent use of *ugly* in the data from Chinese learners.

8.1.5 Further research

The focus of this research is on the use of frequently recurring positive and negative evaluative adjectives and their preferred syntactic and collocational patterning across three varieties of learner (Chinese, French and German) and native-speaker English. A worthwhile replication study would be to examine the same phenomena in those ICLE subcorpora which have recently been compiled, e.g. Bulgarian, Czech, Hungarian, Lithuanian and Romanian.

This study could also be extended to include research on the positioning of these patterns in the interview discourse overall. For example, it was noted that sequences such as *it was good, it's really good* were typically used as a type of summarising evaluative comment on a situation, event or experience. Follow-up research could be conducted on whether other patterns in which

these adjectives are found are used anaphorically or cataphorically. Such a type of analysis would provide further insights into the grammar of speech as a discourse-oriented phenomenon (O'Keeffe et al. 2007). Another avenue for exploration, as suggested by De Cock, is to examine the extent to which preferred patterns in the native-speaker speech are reflected in listening comprehension tasks and input materials in textbooks.

8.2 Comparison of native and non-native speaker learner corpora from a move structure and pragmatic perspective

Summary 8.2

Upton, T. and Connor, U. (2001) Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20: 313–29.

The study examines the politeness strategies used by Americans, Finns and Belgians in a learner corpus of letters of application by advanced students. A key feature of the analysis is that the positive and negative politeness strategies were investigated from a Swalesian (Swales 1990) move structure patterning perspective.

Most research on learner corpora is of a general ELT/EAP nature, as in Case Study 8.1. This case study is one of the few that have investigated learner corpora of an ESP nature. Native-speaker corpora are usually taken as the default value for 'expert' corpora, but in this study a native-speaker corpus was also considered to be within the purview of learner corpora, no doubt for the reason that they were undergraduate university students.

8.2.1 Aims

The main aims of this case study were twofold, one focusing on the methodology and the other on the pragmatic concept of politeness in job application letters:

- to demonstrate the efficacy of a multi-level analysis of a genre-specific learner corpus that included both a hand-tagged moves analysis coupled with a computerised analysis of lexico-grammatical features of text; and
- to show how a pragmatic concept such as politeness can be operationalised to allow for computer-generated counts of linguistic features related to that concept.

8.2.2 Corpora and methodology

The data for this case study were drawn from the Indianapolis Business Learner Corpus (IBLC), created of genre-specific corpora of the written language

of native and non-native speaking students in undergraduate business communication classes (see http://www.liberalarts.iupui.edu/icic/research/indianapolis_business_learner_corpus for more details of this corpus). Specifically, the data consist of a cross-cultural job application simulation among undergraduate university students of business in Belgium, Finland and the USA. There were a total of 153 application letters in the corpus, 70 from Belgium, 26 from Finland and 57 from the USA.

A coding scheme for the corpus was worked out based on the concept of Swalesian (Swales 1990) genre moves categorised by the communicative purpose of individual rhetorical units. Seven moves were identified for the learner letters of application including, as expected, identifying the source of information, applying for the position etc. Upton and Connor noted that two trained raters coded the letters for rhetorical moves with a check carried out on inter-rater reliability. After hand-tagging of the moves, *WordSmith 2.0* was used for generating quantitative data.

Upton and Connor were specifically looking at how positive and negative politeness strategies were operationalised in the corpus. The search queries were based on the linguistic features identified in Maier's 1992 study of cross-cultural politeness strategies, using Brown and Levinson's (1987) model of politeness.

8.2.3 Results and analysis

For the analysis, Upton and Connor decided to concentrate on two moves, Move 4 – indicating a desire for an interview or further contact, and Move 5 – giving thanks for consideration at the end of the letter. At a macro-level of analysis, while Move 4 would no doubt be considered as somewhat of an obligatory move, only 50 per cent of the Belgian letters (35/70) included this move, while 80 per cent of the American letters (45/57) and 73 per cent of the Finnish letters (19/26) did. As for Move 5, it was found that the Americans used formulaic expressions such as *Thank you for … consideration* realising Move 5 far more frequently than the Belgians (40 vs 13 per cent of Belgians), which was even lower for the Finnish students (4 per cent).

As stated previously, the analysis of politeness strategies was based on Brown and Levinson's model. Negative politeness strategies, which are conceptualised as reinforcing the speaker's respect for the addressee, include indirectness often signalled by sentences which begin with words other than 'I', 'you' or 'my', and modals which have the effect of softening the idea being communicated, e.g. *Should you want to discuss my qualifications further…* The Belgian writers used qualifying modals more than twice as often as the Finns (40 to 16 per cent), but about 20 per cent less frequently than the Americans (40 to 51 per cent).

Looking at positive politeness strategies, these are used to emphasise the shared goals and common ground of the speaker and addressee. One aspect,

directness, was realised by sentences that started with the phrase 'You can …' or the phrase 'Please [+ action verb] …' which although polite, give the impression of a command. Such structures, occurring in both Moves 4 and 5, were significantly higher in the letters written by Americans. Optimism, another positive politeness strategy, connects with the addressee's desire to have his or her needs met. This was commonly expressed by 'look forward to' or 'hope', with the Finnish and American writers displaying a higher use of these formulaic expressions in percentage terms than the Belgians.

Upton and Connor conclude from their analysis that it is the *type* of negative and positive strategies used that distinguishes the writers from one another rather than a question of whether they use one strategy or the other. The Americans tended to use more formulaic expressions for both strategies whereas the Belgians used much less frequently employed formulaic expressions and showed more individuality in their letters. The Finns fell between the Belgians and the Americans on this continuum of formulaicity and individuality.

8.2.4 Commentary

The methodology employed in this corpus analysis is noteworthy for its application of a manually tagged corpus for move structure analysis. This is a comparatively small corpus, consisting of just 153 letters, so the tagging by two raters would be manageable. Also, job application letters have a set of move structures which are fairly prototypical of the genre, making it relatively easy to identify the different rhetorical functions. Working with a much larger corpus or a less clearly defined genre would obviously be less straightforward. Another interesting aspect of this analysis is that the linguistic features realising different move structures are investigated from a pragmatic perspective inspired by Brown and Levinson's model of politeness strategies. In contrast to most other studies of learner corpora, the question of native-speaker vs non-native speaker competence as regards grammatical accuracy does not arise with these advanced learners.

In their discussion of the findings, Upton and Connor are suitably cautious, noting that their analyses are only suggestive and merit further verification before conclusions about cross-cultural differences can be made. Why there is such a difference between the American and Belgian writers in terms of style (formulaic vs individualistic) is left open to debate. And do factors other than cultural differences come into play? The background of the Belgian and American participants may offer some insight here. The Belgian and Finnish participants were, on average, younger and less experienced than their American counterparts on the undergraduate business programme, who were returning to university on a part-time basis while continuing to work. The more formulaic patterning for politeness strategies in their letters could perhaps be explained by the fact that they were already acquainted with this genre given that they were in full-time employment and were attending university part-time.

One general point that can be extrapolated from this case study is that there are more often than not unexpected variables and that the analyst has to keep in mind the many variables that should be taken into consideration before drawing conclusions.

8.2.5 Further research

As I indicated earlier, the letters in this case study were drawn from the Indianapolis Business Learner Corpus (IBLC) and were produced by language learners and/or novice writers from the USA, Finland and Belgium. According to Upton and Connor, further additions to the corpus would include letters written by professional native speakers of English as well as by professional and fluent non-native speakers of English. Enlarging the data set with letters written by different cohorts of participants would no doubt reveal nuances in the move structure patterning. At present, very little research exists on learner corpora of ESP from a professional perspective, so the compilation of the IBLC is a very welcome addition to the field (see Section 6.6.2 for accounts of various initiatives on compiling ESP learner corpora in the academy).

Other professional genres which would lend themselves to the kind of move analysis carried out in this study would be fairly conventionalised genres such as various types of business letters. Letters written in English by both native-speaker and non-native speaker expert writers could be move-tagged as in Upton and Connor's study, and then analysed from an intercultural perspective (see Connor 2011). Additionally, business correspondence in other languages could be examined. For example, Lee-Wong (2005) carried out a qualitative analysis of hedging in a corpus of 57 business letters (Taiwan: 10; PRC: 47) comprising a total of 10,452 characters. Hedges were not only used as a politeness strategy, but also found to communicate fuzziness, e.g. *If possible, my company will absolutely not insist as such*, expressing the writer's desire to cooperate while at the same time promising nothing definite (p. 287). Lee-Wong's (ibid.) corpus was not tagged, but examining hedging on a move structure basis would also be another possibility for further research.

8.3 Comparison of expert corpora from an intercultural perspective

Summary 8.3

Mur Dueñas, P. (2009) Logical markers in L1 (Spanish and English) and L2 (English) business research articles. *English Text Construction*, 2 (2): 246–64.

This study investigates the use of metadiscoursal logical markers (additive, contrastive, consecutive) in LI research articles (RAs) in Spanish and English

(*continued*)

and L2 RAs in English in a specific discipline, i.e. business management, to examine whether the use made of these three types of metadiscoursal logical markers by Spanish scholars in their English RAs resembled that in L1 English or Spanish texts.

Intercultural studies of academic genres have been the subject of much research from the 1990s and, as Connor (2004, 2011) points out, corpora are playing a greater role in this field. However, most of these studies have focused on the analysis of L1 academic texts, which have been compared to similar texts in other languages. In contrast, only a few studies, such as this one, make a double comparison of L1 academic genres in the two languages and L2 texts in English.

8.3.1 Aims

The general aim of this research was to explore a specific metadiscourse category (logical markers) in research articles (RAs) from the discipline of business management written in English by Spanish-based scholars. Three types of logical markers were investigated: additive (e.g. *moreover, similarly*), contrastive (e.g. *however, otherwise*) and consecutive (e.g. *thus, accordingly*). These metadiscoursal markers, used to signal not only logico-semantic relationships but also the ensuing interactive reader–writer relationship (Hyland 2005), were compared with those in (a) RAs in Spanish; and (b) RAs in English by Anglo-American authors. The specific aim was to explore the possible influence of the linguistic/cultural context on the frequency and realisations of the three types of logical markers in L1 English and Spanish and L2 business management RAs.

8.3.2 Corpora and methodology

The analysis was based on three subcorpora of business management RAs, forming part of a broader corpus, SERAC (Spanish-English Research Article Corpus), published between 2001 and 2006 (see http://www.interlae.com/ for more publications related to this corpus). The three subcorpora used in this study are as follows:

- ENGBM: comprising 24 RAs written in English by Anglo-American scholars (197,922 words)
- SPENGBM: comprising 24 RAs written in English by Spanish scholars (192,546 words)
- SPBM: comprising 24 RAs written in Spanish by Spanish scholars (166,114 words)

Mur Dueñas (ibid.) reports that articles for the ENGBM and SPENGBM corpora were selected from high impact journals in consultation with specialists who

advised on the most prestigious and commonly read publications in the field. Moreover, the Spanish authors of RAs written in English were contacted to find out about their writing processes. Only those authors whose RAs were originally written in English or had undergone only very minor revisions were included in the corpus. Articles in SPBM were selected from local Spanish journals.

A corpus-driven approach was used to identify relevant markers. That is to say, Mur Dueñas did not start from a benchmark of frequency lists provided by reference grammars for corpus searches or a corpus-derived frequency list. Instead, each text was read and carefully scanned to identify logical markers. Once a particular feature had been identified, all tokens were retrieved using *WordSmith Tools 4.0* and then verified as logical markers through examining the corpus co-text to ensure they functioned as such. In this way, a profile of logical markers was incrementally built up.

8.3.3 Results and analysis

After the counts were normalised per 10,000 words, results revealed that the overall frequency of logical markers was higher in the RAs written in English by Anglo-American authors than in RAs in Spanish. Although it might be expected that the Spanish scholars would transfer the less frequent use of explicit signals to their writing in English, this was found not to be the case in the SPENGBM corpus. Mur Dueñas puts forward several explanations for these findings. Spanish scholars may be aware of the rhetorical differences in terms of the more explicit signalling in the RAs in English, which could be the result of overt teaching of these devices in tutorials, or self-reading of international RAs in English. Another reason offered is that Spanish scholars writing in English may feel the need to be more precise to ensure accurate interpretation by an international audience, regardless of whether they are aware of the rhetorical differences.

As for the types of markers used across the three categories, some general similarities and differences were noted. The range of consecutive logical markers used was similar in the two English subcorpora. A wide range of contrastive markers was also noted in these two subcorpora with the Spanish-English RAs exhibiting a wider range of regularly used contrastive markers. A wide range of additive markers was found in both ENGBM and SPBM; however, this contrasted with the narrow range of markers used by Spanish scholars in their RAs in English. Mur Dueñas explains these divergences in terms of the different way information is presented and developed in the two languages and sociocultural contexts. In the Spanish RAs (SPBM) results and arguments tend to be presented in a cumulative fashion accounting for the higher use of additive markers. In contrast, in the two English subcorpora scholars tend to organise their results by contrasting those which support their hypothesis stated in the introduction with those that do not, and also contrast their arguments to those made in the previous literature or research. Mur Dueñas concludes that the varying discourse-flow pattern of

the international texts in English vs the Spanish ones could be related to the different contexts of publication, i.e. national (local, restricted) vs international (competitive, diverse) setting up a different writer–reader relationship.

In the detailed analysis, it was in the category of contrastive logical markers where the most variation among types of markers was found; *on the other hand, nevertheless*, *nonetheless, conversely*, and *on the contrary* were frequent in the Spanish–English RAs, but infrequent in the English RAs. Mur Dueñas posits that these may be deviant realisations for the signalling of contrast in this genre and discipline. In spite of these deviations, however, Mur Dueñas notes that they did not prevent publication in international journals and would not seem to conflict too much with the expectations of editors, reviewers and readers, falling within acceptable norms. In fact, Mur Dueñas concludes the article by expressing her support for a critical pragmatic approach to scholarly writing, whereby writers should be allowed choices and not have to follow dominant practices.

8.3.4 Commentary

The three subcorpora were compiled according to careful design procedures, with 24 RAs collected for each subset of abstracts. The ethnographic aspect of the data collection is noteworthy in that specialist informants were consulted on the most prestigious and commonly read journal articles in business management and the Spanish authors writing in English were consulted on their writing processes. The corpus-driven method of data extraction, while laudable, must no doubt have been a time-consuming process as the concordance output was scrutinised to ensure that the tokens actually functioned as logical markers in the co-text. In more qualitatively oriented studies of this kind, the issue of balancing sufficient data to ensure representativeness against the amount of data that can realistically be analysed often arises, and is another question to which there is no easy answer. As this study involved investigation of a specialised discipline in one particular genre, 24 reports would seem to be sufficient for making generalisations about the data. And, indeed, as Biber (1990) has pointed out, small corpora of the size in this study are perfectly adequate for investigation of relatively common grammatical items.

As the number of words differed across the three subcorpora each consisting of 24 reports, results were normalised per 10,000 words for the overall count. However, the more detailed comparison of individual types for each of the three kinds of logical connectors is made using the number of tokens without any other kind of statistical processing. Although the two subcorpora of English RAs were of a comparable size, the corpus of RAs in Spanish comprised 166,000 words as against around 195,000 words. In corpus comparisons achieving balance across the number of texts and tokens presents difficulties. However, Oakey (2009) does not see this as a problem and argues in favour of isotextual comparisons in which comparisons are made across similar numbers

of communicative acts rather than across similar amounts of language, i.e. isolexical comparisons. Mur Dueñas's comparison could thus be viewed as isotextual as balance is maintained through the same number of reports across the three corpora, although the number of tokens differs.

Like De Cock (see Section 8.1), Mur Dueñas puts forward a number of explanations for the data, which relate to the different audience and contexts of publication, or to the background of the journal authors.

8.3.5 Further research

The ethnographic dimension of the study has been remarked on in the data collection stage and this could be extended to other stages of the study. Mur Dueñas herself advocates that corpus-based analyses should be informed by specialists and gatekeepers such as reviewers and editors to shed light on the writing process and to inform to what extent transfer or deviant use of L1 rhetorical features may prevent publication. Journal article writers could also be interviewed for their views on the use of metadiscoursal features.

This research has concentrated on a double comparison of RAs in English and Spanish in a particular discipline and genre. As genres are subject to variation and different disciplines have their own discursive practices (Hyland 2005), similar corpus-based intercultural analyses could be conducted to determine whether the transfer process is absent, as in this study, or present, in other types of scholarly writing and also across other languages.

Mur Dueñas's study is restricted to logical markers. The logico-semantic relations of addition, contrast and consequence examined in this article can also be achieved through other lexico-grammatical choices. One suggestion for a follow-up study is to examine other linguistic devices realising a particular relation. For example, addition can also be signalled via *-ing* complement clauses. Moreover, the study concentrates only on those markers which join two main clauses and thus have a metadiscourse function. The study could be modified to examine how specific markers function, not only globally but at a more local level of coherence; for example, *thus* and *as a result* can also be used intrasententially and it would be enlightening to observe the global and more local uses of such markers from an intercultural perspective.

8.4 Investigation of collocational behaviour from a social psychological perspective

Summary 8.4

Pearce, M. (2008) Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora*, 3 (1): 1–29.

(*continued*)

In this case study Pearce examines how men and women are represented in the 100-million-word BNC using the *Sketch Engine* corpus query tool to examine the collocational and structural behaviour of the noun lemmas MAN and WOMAN (i.e. *man/men* and *woman/women*). The corpus data are analysed from a social psychological perspective.

There has been growing interest among corpus linguists as to how certain groups of people are represented in the news (see Section 4.3.3 for a discussion of Baker and McEnery's work on how refugees and asylum seekers are represented in UK broadsheet and tabloid newspapers). Pearce maintains that examining the behaviour of the lemmas MAN and WOMAN in the general 100-million-word BNC would reflect how the roles of men and women are perceived in society at large.

8.4.1 Aims

The aim of the research was to examine how men and women are represented in the BNC, analysed from a social psychological perspective according to:

- how they are described and categorised;
- how they are represented as 'doing and experiencing' as agents;
- how they are represented as 'undergoing' as patients and beneficiaries.

8.4.2 Corpora and methodology

In other studies reported in this volume, the 100-million-word BNC (http://www.natcorp.ox.ac.uk/) has been used as a reference corpus for benchmarking the salient lexis or keywords of the smaller corpus under investigation. In this study the BNC is used as the object of enquiry for researching the collocational behaviour of MAN and WOMAN. While gendered items have already been researched in other corpus-based studies, Pearce's study is noteworthy for his use of the *Sketch Engine* tools, which, he states, enable a more sophisticated analysis of the social and cultural meanings these items entail on account of the search facilities afforded by this software. A more detailed collocational picture of the target lemma can be built up through the production of a 'word sketch' than through traditionally derived collocates, as shown below.

Concept 8.1 Sketch Engine search facilities

Pearce explains that:

> Table 8.1 [Table 8.3 below] shows the top ten items collocating with MAN in three grammatical relations. G1 is the relation between a verb and

its subject (e.g. *The man died*). G2 is the relation between a verb and its object (e.g. *The officers arrested the man*). G3 is the relationship between an attributive adjective and the noun it modifies (e.g. *The mourners were young men*).

The first set of numbers in the second row refers to the number of times MAN appears as the subject of a verb (19,174 times), object of a verb (15,847 times) and premodified by an adjective (28,802 times). The second set of figures in the second row is a calculation of the likelihood of MAN occurring in these relationships, as subject, object or with adjective modifier, compared with nouns in general, i.e. MAN is 4.0 times more likely to occur as the subject of a verb than nouns in general.

Sketch Engine also gives a list of the lemmas which have statistically significant occurrences with the target lemma. For example, in row 3 of the table, *die* is the most significant collocate for MAN with a saliency of 31.84 (even though it occurs less frequently (275 occurrences) than *stand* (331 occurrences).

Table 8.3 Part of a word sketch showing three grammatical relations for MAN (Pearce 2008: 4)

G1	***Subject***		***G2***	***Object***		***G3***	***Adjective modifier***	
	19,174	***4.0***		***15,847***	***1.7***		***28,802***	***2.4***
die	275	31.84	arrest	225	38.42	young	3,719	65.69
stand	331	29.95	kill	318	34.73	old	2,431	51.26
sit	274	29.26	accuse	132	28.86	gay	205	43.08
walk	197	28.05	convict	64	28.54	tall	355	42.95
wear	193	27.57	marry	108	28.52	middle-aged	138	41.56
live	212	24.86	age	65	28.47	older	352	39.03
come	619	25.53	jail	55	28.12	wiser	160	37.78
work	267	22.46	charge	123	26.29	homosexual	93	37.62
look	341	21.65	meet	266	26.27	younger	224	37.05
nod	56	21.51	name	102	25.32	married	209	39.91

In addition to showing which grammatical roles a lemma prefers or avoids, another search facility, *Word Sketch Difference*, allows two target lemmas to be compared. For example, in the BNC, MAN and WOMAN both occur as subject of the verb *scream*, although *scream* is more strongly associated with women as revealed by its saliency profile; i.e. a saliency of 18.6 for WOMAN as opposed to 8.6 for MAN.

(*continued*)

Another function of this tool is that it identifies combinations which occur exclusively with one lemma in the target pair. For example, WOMAN, but not man, is modified by the adjectives *pretty, dumpy* and *scarlet*, whereas MAN, but not WOMAN, is modified by *burly, balding* and *dirty*.

8.4.3 Results and analysis

Using the *Sketch Engine* tools outlined in Concept 8.1 above to extract collocational and structural profiles of MAN and WOMAN, Pearce discusses the data from the perspective of five broad categories: power and deviance, social categories, personality and mental capacity, appearance and sexuality, three of which are summarised below.

As noted by Pearce, the BNC data revealed important asymmetries in the way men and women are represented in relation to power and deviance. Verbs collocating strongly with MAN as subject relate to the exercise, or ownership, of power. For example, *dominate* and *lead* associate more strongly with MAN than WOMAN as do the verbs *possess* and *own*. Pearce concluded that such patterns reflect contemporary disparities of wealth, power and resources, which is also reflected by the type of attributive adjectives (*great*, *influential* and *leading*) collocating strongly with MAN. Pearce also noted the association of MAN with attributive adjectives (e.g. *cruel, dangerous, convicted, evil-looking*) denoting involvement in criminal and deviant activities, commenting that this was unsurprising given that UK Home Office government statistics show that 85–95 per cent of crime is committed by males. By contrast, WOMAN strongly collocated with adjectives signalling various social categories such as marital/reproductive status, nationality and ethnicity.

The BNC data have also shown there to be differences in the mental and behavioural characteristics of men and women. Pearce analysed the data for this third category, within a taxonomy of human personality based on the 'lexical hypothesis', which states that language encodes the ways in which people differ and that the *accumulations* (my italics) of such descriptors can lead social psychologists towards pinpointing individual differences. At the broadest level of abstraction five dimensions have been proposed: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience (Goldberg 1981, cited in Pearce 2008).

Pearce used the *Word Sketch Difference* tool to extract collocational data for the adjectives (attributive and predicative) realising these five dimensions, namely those shared by MAN and WOMAN and those occurring exclusively with the target lemma. An interesting aspect of Pearce's study is that he compared the corpus data with findings from social psychology research into gender and personality (Costa et al. 2001, cited in Pearce 2008). In general, the BNC corpus

data represented gender in stereotypical ways. As Pearce points out, this is not surprising given that gender is a social construct, established and reproduced in discourse. The research from the literature that men generally test higher than women in facets of extraversion and openness to experience was borne out by the corpus data; for instance, the adjectives *astute, scholarly*, and *self-educated* were found to occur exclusively with MAN. Women were portrayed as neurotic (e.g. *distraught, hysterical* and *silly*). It has to be noted that these results are from the whole BNC. When Pearce examined a few of these collocations for WOMAN in more detail he found that they mainly occurred in prose fiction texts, even though this text-type constituted only 16 per cent of the BNC. Pearce thus notes the presence of a 'sociological discourse' in which women are presented as objects of enquiry in stereotypical roles in this text domain.

Along the agreeableness and conscientiousness dimensions MAN ranked consistently higher than WOMAN with *affable, amiable-looking*, *avuncular, good-natured, jovial* and *personable*, and *braver*, *humane, law-worthy, patient, sincere* and *tolerant* among the adjectives exclusive to describing MAN for the agreeableness and conscientiousness dimensions, respectively. However, these data contradict the social psychology literature which recounts that in most cultures women are regarded as more dutiful than men (Costa et al. ibid.).

8.4.4 Commentary

This case study deserves comment for the tools utilised and social psychological perspective for interpretation of the corpus data. As noted in Section 4.1, there has been much debate on corpus-based vs corpus-driven approaches. Those proponents in the corpus-driven camp advocate approaching the data as a *tabula rasa* without any annotation, to allow new hitherto unexplored aspects of language to emerge. The *Sketch Engine* tools would seem to marry the best of both approaches. The POS and parsing functions with their paradigmatic orientation serve as an initial filter for classifying the lexis according to grammatical relations. The further processing of the data for statistically significant collocates as shown in Concept 8.1 makes for a more syntagmatic analysis. This combinatorial approach presents the data in a much more systematic and accessible manner from which valuable syntagmatic information can be read off, such as semantic preferences and prosodies.

Interpretation of raw corpus data can be problematic especially when socio-cultural issues are involved such as in this study. Pearce adopts an eclectic approach to interpretation. For example, he cites government statistics to corroborate the finding that a high proportion of adjectives signifying criminal activities co-occur with MAN. He explains the patterns of differences with respect to personality traits through alignment with research findings on the psychology of gender. While some corpus data on personality traits, e.g. extraversion, align with the research literature, other data such as those

for agreeableness and conscientiousness do not, but Pearce does not offer any other explanations for this. The multi-method approach proposed by Sealey (2009), using insights from realist social theory, complexity theory combined with corpus linguistic findings, may offer some help here (see Section 5.4.1). But to adopt Sealey's approach it would be necessary to have much more sociological data, which is not encoded in the BNC. It is also a matter of importance to interpret corpus findings in the light of the nature, size and period of the corpus. As the BNC contains texts from 1975–94, the stereotypical portrayal of gender differences may not be so surprising.

8.4.5 Further research

This study shows the role that sociolinguistic data can play in the interpretation of corpus findings. Pearce, himself, has suggested building up a more nuanced picture of gendered language through consideration of a wider range of grammatical relations and examination of other gendered binaries (e.g. *boy/girl, male/female*).

Another avenue for further exploration would be to explore the collocational patterning of the items in a more recent and larger corpus than the BNC, such as the 450-million-word Bank of English (http://www.titania.bham.ac.uk/). A diachronic analysis with a view to charting how perceptions change from one period to another would be another suggestion for further exploration. The 400-million-word Corpus of Historical American English (1810–2009) would lend itself to such a study (see http://corpus.byu.edu/coha/). For example, in a brief overview of this corpus on the aforementioned webpage, Mark Davies notes that concordance output of the changes in collocates of nouns such as *woman* from different time periods shows cultural shifts over time.

Pearce's study could be extended to take account of the text types in which certain collocates appeared, which may well have a bearing on the conclusions reached. Alternatively, a similar study could be restricted to a particular text type, such as Caldas-Coulthard's (1993) study, which examined differences in the ways in which what women and men in positions of authority had said were represented in written news.

8.5 Comparison of media corpora from a discourse-analytic perspective

Summary 8.5

Duguid, A. (2009) Loud signatures. Comparing evaluative discourse styles. In U. Römer and R. Schulze (eds) *Exploring the Lexis–Grammar Interface*, pp. 289–315. Amsterdam: John Benjamins.

This corpus-assisted discourse-analytic study looks at evaluative discourse styles in two small corpora of opinion articles from British broadsheets and the *Times Literary Supplement*. The study is of interest as it combines qualitative and quantitative methodologies and makes use of three different types of software, *WordSmith Tools*, *Concgram*, and *WMatrix* to identify the salient resources for conveying evaluation.

This study belongs to part of a wider research paradigm known as Corpus-Assisted Discourse Studies (CADS), which is critically oriented following Fairclough's approach to CDA (see Section 4.3.3). Such studies often use a Hallidayan approach to analysis, as is the case in this study, which borrows from Martin and White's (2005) complex system of Appraisal for its conception of evaluative language. The aspect of the Appraisal system of relevance here is the commitment to the appraisal expressed and how this can be reinforced or downplayed, characterised as 'loud signatures' (pp. 203–6), which can involve intensification, quantification and hyperbole, often employed for ironic effect.

8.5.1 Aims

The general aim of this research was to compare evaluative styles in two small corpora of opinion articles from British broadsheets and the *Times Literary Supplement*. Duguid's main research question was to see whether it was possible 'to identify the patterns or repeated regularities of forms in the ways in which loud signature texts present their evaluations of experiences through language' (p. 295). In so doing, her specific aim was to show how press columnists flout the typical priming effects found in press reportage, thereby exploiting the expectations of the reader (see Concept 1.11).

8.5.2 Corpora and methodology

Two corpora were compiled for this study: a humorous opinion pieces corpus (HO Corpus) made up of articles reviewing a variety of artefacts and events, consisting of 256,353 tokens and over 200 texts from the years 2004 and 2005, and a comparable corpus, made up of reviews published by the *Times Literary Supplement* (TLS Corpus), also from the same years, totalling 216,650 tokens. Duguid describes the TLS Corpus as a 'background' one as the focus of her study is to foreground the findings from the HO Corpus.

The quantitative methodology involved making frequency lists and running a keyword comparison. This was followed up by a more finely grained qualitative analysis using *WordSmith*, *ConcGram* and *WMatrix* to tease out various aspects of the key resources for building evaluation.

8.5.3 Results and analysis

The qualitative analysis, the main focus of the research, spans several areas, namely the writer/reader relationship, deictic elements, loudness and figurative language covering intensification, hyperbole and irony. Keyword lists revealed greater emphasis on dialogistic pronouns in the HO Corpus than in the TSL Corpus, which project a conversational tone linking reader and writer. Duguid then used *ConcGram*, which allows for constituency and positional variation (Greaves 2009), to extract phrases with *I* and *you*, e.g. *And **I'm** also being realistic when **I** tell **you** that in a straight fight*…(see Example 1.5). Following Martin and White (ibid.), examples such as this construe the writer as maximally committed to the value position being advanced and its projected evaluation, with the dialogistic pronouns strongly aligning the reader to that said value system.

Another key difference between the two corpora, as revealed by the concordance output using *WordSmith Tools*, was in the use of deictic elements involving demonstrative reference. Whereas these were endophoric in the TLS Corpus, there was a large proportion of these used exophorically in the HO Corpus, e.g. *She wears one of **those** blue suits that only women MPs seem to be able to find*. Duguid explained that the concordance lines show how this usage creates a presupposition of shared experience and values: 'We might say then that the reader is being asked to recognize the experience to which the writer is referring, in order to understand the quality of evaluation' (p. 299), which may involve the notion of stereotypes. In such a way, evaluations are embedded inside noun phrases containing pre- and post-modification of the head noun.

Duguid also commented on register differences between the two corpora signalled by features of 'loudness'. To retrieve such types of 'loud signatures' typical of informal conversations, Duguid used the POS tagger function from *WMatrix*. A comparison of the HO Corpus with the BNC Spoken showed a number of items linked to what Martin and White term 'force', involving the use of intensification and quantification, upscaling and maximisation. Superlative adjectives and superlative adverbs of degree reflecting intensification were found to be salient in the HO Corpus. Hyperbole in the HO Corpus was achieved through lexical build-up, repetition and intensification, e.g. *But it still felt brilliant. **Utterly, stunningly, mind blowingly, jaw droppingly** brilliant.*

8.5.4 Commentary

This is an interesting study on several accounts. While taking a discourse-analytic approach to the evaluative discourse of opinion articles, this study could also be seen as bordering on a corpus-stylistic analysis uncovering aspects of creativity in a genre which has some affinity with the type of bantering found in casual conversation. The initial keyword and more detailed lexico-grammatical analyses have indicated the inherent literariness of humorous opinion pieces through uncovering such phenomenon as hyperbole, etc.

The methodologies adopted also deserve some comment. In common with many other studies, this research starts from a quantitative analysis through extraction of frequency and keyword lists. However, whereas most other studies use *WordSmith* to examine the lexico-grammatical patternings of the keywords, this study, in addition, makes use of *WMatrix* and *ConcGram* for POS tagging and for revealing positional and constituency variation in the lexico-grammatical patterning. One point of interest is that Duguid used the BNC spoken component (approximately 10 million words) as the reference corpus for benchmarking the POS frequency lists from the HO Corpus. No doubt, this would be for the reason that she considers the HO Corpus to be more akin to spoken language with its conversational, informal overtones.

8.5.5 Further research

Although this study made use of *WMatrix*, only one of the facilities was utilised, i.e. the POS tagger. The semantic analysis tagger could also profitably be used to delineate the topics of the humorous opinion pieces in broadsheets vs the more measured and serious reviews in the TLS Corpus. It is to be noted that compilation of the corpora spanned the same time period, i.e. the years 2004 and 2005. A more diachronic perspective on the data could be incorporated through alignment of the topics with the date of publication. Peaks and troughs of certain topics could be plotted to see if they corresponded to key events in politics, economics etc., similar to the type of analysis carried out in Gabrielatos and Baker's study on refugees and asylum seekers (see Section 4.3.3).

8.6 Investigation of lexico-grammar constituting discursive practices in an international workplace setting

Summary 8.6

Handford, M. and Matous, P. (2011) Lexicogrammar in the international construction industry: A corpus-based study of Japanese-Hong-Kongese on-site interactions. *English for Specific Purposes*, 30 (2): 87–100.

This article reports a situation which reflects how globalisation is increasing opportunities for NNS-NNS interactions. The research is part of an interdisciplinary study based in the Department of Civil Engineering at the University of Tokyo. The study is notable for its use of multiple data sources and ethnographic dimension to understand how the discursive practices of NNS engineers in their daily working life on-site are construed through the lexico-grammar.

A few years ago, most corpora of oral discourse, both in the academy and professional workplace, were analysed without recourse to videotapes of the interactions.

Moreover, to date most studies of professional oral discourse have involved native-speaker interactions. This study advances the field as it is situated within the wider paradigm of ELF communication (see Section 6.1). Importantly, it also makes use of visual and non-verbal data (see Section 4.5), thus extending Norris's (2004) methodological framework for analysing multimodal interactions to corpus work, a paradigm which is now being taken up by discourse analysts working in multimodality (see, for example, Baldry and O'Halloran, in press 2012).

8.6.1 Aims

Handford and Matous frame the research questions as follows:

> How do the English interactions recorded in a Japanese company in Hong Kong compare at the lexicogrammatical level to everyday English and business English?
>
> How can such lexicogrammatical items be interpreted to shed light on the context they reflexively constitute?

8.6.2 Corpora and methodology

This study makes us of several data sources collected over several months including audio and video recordings, interviews, expert informant insights and observation notes from a construction project involving a Japanese/Hong-Kongese/European joint venture in Hong Kong. The methodology for this ethnographic case study involved the following stages:

Step 1: collect and transcribe relevant textual and ethnographic data.
Step 2: pinpoint potentially important lexico-grammatical items and linguistic/paralinguistic features.
Step 3: understand how the situated meaning, the chosen item or the feature is invoked in its specific context.
Step 4: infer the practices, goals, socially situated identities and social structures that orient the participants through the discourse in which the item occurs.

The authors point out that while their methodological approach prioritises the on-site interactions, at the same time, the analysis is cyclical in that 'it moves between the text and the context to understand the meanings, goals, practices, identities and structures the participants have constructed through their talk', exemplified as follows:

Social practice:	Managing projects
Professional practice:	Ensuring resources are used appropriately
Discursive practice:	Checking
Text:	*No oil + oil is empty*?

The focus of this paper is primarily on quantitative data analysis and concerned with the interpersonal aspects of communication in international spoken discourse in the construction industry, examining, for example, how pronouns are used for signalling social relationships, backchannels for listener solidarity and deontic modality for negotiating power over actions.

8.6.3 Results and analysis

The authors first conducted keyword and cluster searches. Comparisons between this HK data and SOCINT (a corpus of social and intimate (everyday) conversation and CANBEC (Cambridge and Nottingham Business English Corpus) (see Section 6.2), were made to provide answers to the first research question noted above.

Keyword results revealed that transactional 'construction' words were statistically significant, which the authors classified into six key types of engineering construction nouns, e.g. site nouns (*slope*), process nouns (*excavation*). Of note is that almost half the most significant items in the HK data were either interpersonal (e.g. backchannels *hmm)* or discourse markers (e.g. *so*). Deictic markers, e.g. *this is*, *this one is*, were a key item in the two- and three-word clusters, supported by the video recordings which showed participants constantly gesturing. Thus clarification of meaning was achieved through combined verbal and physical deixis. A high degree of verbal and non-verbal deixis was also noted for clarifying explanations while drawing plans, e.g. *And then and (I go to this here) they can connect + here*….

Deontic modality was also specifically investigated to examine how the interlocutors negotiated what actions needed to be taken. That certain items such as *we* and *have to* and *need to* co-occurring with hedging expressions such as *I think* were preferred to more direct expressions suggests that the interlocutors were aware of face needs. However, at the same time, the ethnographic data revealed some cultural dissonances, such as the fact that the Japanese approached communication with 'a more hierarchically stratified mindset than the Hong-Kongese' and there was some confusion over an individual vs. an organisation's role of responsibility on account of minimal written contracts in Japanese companies, aspects which created some fractious decision-making. This observation concurs with previous research on culture and professional groups in which policies and procedures associated with particular professional cultures were found to transcend the borders of organisational cultures (see Spencer-Oatey and Franklin 2009).

On the other hand, the comparison of keywords and clusters with SOCINT and CANBEC revealed similarities in interpersonal language, thus suggesting that the same discursive practices are being applied across separate institutional settings and that equivalent interpersonal concerns are being addressed. Another area of similarity between the HK data and the two reference corpora was the prevalence of problem-solving words, e.g. *make a decision*, indicating that problem-solving is a key workplace skill. As a point of reference Poncini's

(2004) research on discursive strategies in multicultural business settings suggests that intercultural differences (cf. Korean/HK) are much less significant than differences in terms of institutional cultures which, when the same, have the effect of neutralising interethnic differences, which would seem to reflect Handford and Mateus's findings.

8.6.4 Commentary

This is an exemplary small-scale study which moves from description, to interpretation and finally towards evaluation, which as Candlin (2002) states, is a key aim of professional discourse analysis. The study also illustrates how corpus data are but one piece of the puzzle in analysing professional discourse and underscores the value of integrating and comparing corpus data with other data sources and data sets. Although the transcription totals just over 12,000 words of interaction, it has to be borne in mind that this is supplemented with ethnographic data from a range of other sources and that the nature of on-site interactions necessitates that they are brief. No strong claims are made by the authors about the generalisability of the findings; nevertheless, they have partially addressed this issue through comparison with SOCINT and CANBEC. This joint project between researchers from the Departments of English and Civil Engineering at the University of Tokyo leads the way for other collaborative ventures of an interdisciplinary nature. The findings also have important implications for training in intercultural professional communication with regard to problem-solving, summarising and tactical interrupting.

8.6.5 Further research

This study points the way for future corpus-based projects which are anchored in a rich ethnographic context for examination of discursive practices. This research also shows that small-scale multimodal corpus-based analyses are feasible; a hand-held camera was used and the analysis did not require the substantial resources allocated to those projects reported in Section 4.5.

The focus of this research is on quantitative data, but more qualitative analyses could also be conducted to look at extended exchanges involving turn-taking. While not discussed at length, the authors acknowledge the role that English as a lingua franca plays in intercultural professional communication (cf. Bargiela-Chiappini et al. 2007; Connor 2011). Of interest is that the researchers note that daily onsite interactions were successfully achieved and enabled through the use of diagrams, photos, gestures and a high level of shared engineering knowledge, in spite of constrained language proficiency in terms of article usage, word order, negation, tenses, ellipsis and subject–verb agreement. However, Handford and Mateus do not elaborate on this point.

In Section 6.1 the ethnographic dimension of Mauranen et al.'s (2010) Studying English as a Lingua Franca project was noted, one aspect of which

was to look at adaptive processes in action. One suggestion for further research, therefore, is to examine in a corpus-based multimodal analysis not only paralinguistic features such as gestures, but how other support artefacts in the form of diagrams and notes influence the ongoing discourse and can aid successful ELF communication.

8.7 Analysis of a corpus of adolescent e-mails in health communication

Summary 8.7

Harvey, K. et al. (2008) Health communication and adolescents: what do their e-mails tell us? *Family Practice*, 25 (4): 304–11.

In this research Harvey et al. analyse a corpus of e-mails sent to an adolescent health website. The value of the study lies not only in the insights it provides into how adolescents frame their health problems, but also in the issues it raises for health care education and advice.

The increasingly important role that digital technologies are now playing in shaping and changing current communicative practices cannot be underestimated. Today's 'internet generation' of adolescents has grown up in a world mediated by such technologies. This study contributes to the growing body of research on such hybridised corpora, which, while in the written mode, display many features of spoken language (see Section 4.6).

8.7.1 Aims

The main aim of this study was to investigate the concerns and difficulties relating to communication among adolescents seeking online health advice.

8.7.2 Corpora and methodology

The authors analysed the content of a 1-million-word corpus of e-mails sent to the adolescent health website, Teenage Health Freak (http://www.teenagehealthfreak.org). This interactive site, launched in 2000, is designed to provide confidential and evidence-based advice and information on a broad range of health issues to adolescents. The corpus for this study consists of 62,794 messages sent to the website between January 2004 and December 2005.

The methodology consisted of keyword, collocational and concordance analyses. *WordSmith Tools 4.0* was first used to create a list of keywords (i.e. words that are unusually frequent in comparison with general everyday English) as a means of determining the language variety characterising the health language of adolescents. The reference corpus used in this study was the 1-million-word collection of general spoken English from the British National Corpus on the

grounds that a spoken rather than a written corpus is a more appropriate comparator for the informal register of the e-mail messages.

Having derived a list of keywords, the researchers then conducted a collocational analysis to explore the meanings of the keywords in greater detail. Collocations were identified by statistical measures using an MI (mutual information) rating of >3, taken as being indicative of a strong collocation (see Section 1.2.2. on statistical vs. textual collocation). While a collocational analysis provides a general overview of the topics and themes of the keywords, it gives limited information on how words function in context. A concordance analysis was therefore conducted to gain a more detailed understanding of the keywords and their collocations. The large number of concordance lines made it impossible, or infinitely time-consuming, to examine each one in detail (the authors report that there were 2818 instances of the keyword 'tell' alone). The authors therefore decided to adopt the procedure advocated by Sinclair for analysing a large number of occurrences. They first randomly selected 30 concordance lines, and after analysing the patterns, proceeded to examine another 30 randomly chosen concordance lines. This procedure continued until a saturation point was reached, where no new patterns were obvious (cf. Baker 2006).

8.7.3 Results and analysis

The initial keyword analysis showed that, when compared with general spoken English, the e-mail messages contained a number of keywords pertaining to communication, specifically verbs and nouns associated with advice seeking, e.g. *ask*, *question*, *advice*. Of interest is that the follow-up collocational analysis presented evidence showing that negatively laden collocates surrounded the keywords. For example, communication keywords such as 'tell', talk', 'ask' and 'answer' were found to occur with adjectives describing the teenagers as being 'afraid', 'scared' and 'worried'. The more detailed concordance analysis broadened the picture by showing how doctors, parents, family and friends were seen as problematic sources of health advice, as illustrated below.

Results 8.2

Although adolescents reported a strong desire to discuss health concerns with others and obtain professional and familial advice, they displayed reluctance to confide in others, and were uncertain whether it would be right to disclose problems to other people: e.g.

> *talk* to my parents or sisters about it, im too embarrassed. what should i do?
> Do I need to *tell* my mum and dad im an alcoholic at 15?

> Furthermore, extended concordance lines of 'GP' and 'parents' revealed reasons for the non-disclosure of troubles, such as doubts about whether parents would be able to provide appropriate help and support or whether there would be potential rebukes.
>
> (Adapted from Harvey et al. 2008: 308)

8.7.4 Commentary

This research on analysing health communication is unique in that it is the first study to explore a corpus of e-mail texts from adolescents concerning their attitudes to communication. This research also differs from previous analyses on health care communication in two main respects. The study seeks to explore what adolescents themselves see as pertinent issues rather than adopting an 'outsider perspective' in which the issues for research are based on what the researchers deem to be important. Whereas other research on health care communication has examined participants' responses to professionally initiated actions, the focus of this research is on patient-initiated actions. Moreover the findings from this research are not purely descriptive as the intention of the researchers is to make their findings available to health care providers and users of health care services in the form of a practical, encyclopaedic resource, thereby providing a valuable contribution to the lifelong learning and development of different user groups in the NHS.

8.7.5 Further research

The above analysis could be refined to examine the data from a more sociolinguistic and diachronic perspective. The corpus was first set up in 2000 and has the status of a 'monitor' corpus (cf. Sinclair 1991). The corpus could be divided into different time-based subcorpora to facilitate longitudinal analyses with a view to examining how the participants' concerns and dilemmas, and possibly the register have evolved over a ten-year period.

An analysis of the data taking into account sociolinguistic variables of gender and age would no doubt reveal differences which may be of use for educational purposes. However, a major obstacle hindering corpus research on e-mail communication is privacy issues. The authors of the present study note that the Teenage Health Website possesses a privacy policy informing contributors that their requests may be used for research purposes and that participants consent to the collection and use of data which they provide. One dilemma for corpus researchers, though, is how many personal details participants would feel comfortable in revealing for a more fine-tuned sociolinguistically motivated analysis.

The methodological procedures adopted in this study could be applied to other forms of e-mail communication, such as student queries to teaching

staff. In this case, a predicted more formal register would necessitate a different reference corpus for benchmarking the keywords. In fact, the choice of a suitable reference corpus is not unproblematic and it may be advisable to experiment with using different reference corpora to see how they affect the results of the comparative keyword analysis (see Scott and Tribble 2006 for further discussion on this issue).

8.8 Investigation on the effectiveness of corpus-based vs traditional teaching materials

Summary 8.8

Boulton, A. (2010) Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60 (3): 534–72.

This paper reports the results of an experiment with lower-level learners to examine the effectiveness of corpus-based materials and a DDL approach (Johns 1991) compared to more traditional teaching materials and practices. Pre- and post-tests showed both were effective compared to control items and student questionnaire responses were more favourable to DDL.

The lack of empirical evidence for the effectiveness of data-driven learning has been remarked on in Section 7.3.3. Moreover, the few studies which have been carried out tend to evaluate students' opinions rather than their performance. Boulton's paper serves to redress this imbalance through a quantitative study using pre- and post-tests. One significant aspect of Boulton's study is that, unlike other studies of an evaluative nature which survey hands-on concordancing activities, Boulton makes a case for using paper-based corpus materials with novice learners. He also argues that such materials are not incompatible with DDL as they can be used proactively as well as in a more traditional, teacher-controlled way.

8.8.1 Aims

The main aim of this experimental study was to see how lower-level learners cope with paper-based corpus materials and a DDL approach compared to more traditional teaching materials and practices.

8.8.2 Corpora and methodology

The participants in this study were second-year intermediate students enrolled at an architectural school in the north-east of France. None of the 62 students who completed the pre- and post-tests had any prior experience of DDL. The 15 problematic language items to be tested were selected from the learners' own production of an argumentative essay on a topic of their choice loosely

related to architecture. Problem areas were selected on the basis of frequency of occurrence, generalisability and on grammar/usage items which lend themselves to a DDL approach. A third of the 15 items featured in the pre- and post-tests as a control while the remaining 10 were included in the experiment. For students in half of the groups, items 1–5 were taught using traditional methods and items 6–10 using the DDL materials. This situation was reversed for students in the other group. In such a way, no groups or language item received special treatment and served as a control for the others.

The corpus-based paper teaching materials were produced from the 100-million-word BNC using Mark Davies's BYU interface (http://corpus.byu.edu/bnc/). The concordance output for the experimental data consisted of around 5–30 lines for each specific query. The questions were preceded by a short introduction in French to corpora and their potential applications, together with examples. For these DDL materials, students were encouraged to work in pairs or small groups with minimal teacher intervention. The traditional teaching materials were based on selected entries from monolingual or bilingual online dictionaries. In these instructional sessions, the teachers were asked to adopt a comparatively traditional 'knowledge transmission' paradigm prevalent in France. At the end of these experimental sessions, 71 students completed a short questionnaire in French surveying their opinions on both types of materials and practices.

The pre- and post-tests consisted of 30 questions each, with two questions on each of the 15 language items in a multiple-choice gap-fill format, which was familiar to students.

8.8.3 Results and analysis

Using a battery of statistical tests (one-way and two-way ANOVA) Boulton compared the data in two main ways: changes between the two tests and differences between the treatments. The main finding was that scores improved significantly between tests following both traditional and DDL treatments. This improvement in scores could not be attributed to a test effect, as the five untreated control items did not improve significantly. The second main finding was that overall DDL treatment was more effective than the traditional treatment at an 85 per cent level of confidence. Moreover, scores improved more under the DDL treatment for seven of the ten language items, and more learners increased their scores for the DDL treatment than for the traditional treatment. Of note, as Boulton mentions, was the wide range of scores for the DDL items, suggesting that the experiment concealed considerable variation and that learners had different responses to the approach.

Questionnaire responses revealed that students were generally favourable to both approaches with some students commenting that they viewed them as complementary. One interesting finding was that although 51 out of the 71

students who answered the questionnaire said that they would like to continue with corpus data and DDL activities, Pearson's correlation coefficient showed there to be a slight but not significant *negative* correlation between participants' level, as measured by their start-of-year TOEIC scores, and their strength of response to this question (not agree to strongly agree on a 5-point LIKERT scale). It could be surmised that students who had learnt successfully via traditional teaching methods in the past might well like to continue in this way in future.

8.8.4 Commentary

This case study is an important contribution to the field as it represents one of the few studies which provide empirical data on the effectiveness of DDL. Its strength lies in the robustness of the statistical analyses and the fact that the findings provide firm empirical data in support of corpus-based DDL materials. Other quantitative-based studies have surveyed hands-on concordancing; this paper opens up a new line of enquiry with its focus on paper-based materials while still maintaining an essentially DDL approach. Another issue that this paper confronts is the commonly held assumption that DDL is most suitable for use with more advanced learners. Boulton's study has shown, though, that intermediate-level learners can derive benefit from DDL materials. These findings thus support Boulton's proposal that an ideal way for introducing lower-level students to DDL would be from paper exercises to hands-on activities and from pre-set to more open-ended explorations.

In spite of the mostly favourable findings towards DDL, Boulton cautions that one cannot make sweeping generalisations as to its efficacy on the basis of just one single experiment. He also points out that the experiment did not gauge the 'incidental' learning of other language points not explicitly taught that can occur in DDL activities.

8.8.5 Further research

This case study is a 'must read' for practitioners who are thinking of undertaking empirical research of a similar kind, not only for its exemplary methodology, but also for its in-depth review of previous empirically-based studies.

This study also prompts other unexplored directions for future research. One would be a comparison of the benefits of paper-based materials against hands-on DDL work, as suggested by Boulton. Other avenues for enquiry would be to explore in more depth the effectiveness of DDL with different levels of learners, and with students from different backgrounds (e.g. arts vs engineering). Based on the test and questionnaire findings, strongly suggesting that different learners react to the DDL approach very differently, Boulton poses the question 'What type of learner takes to DDL most readily, and is it possible to provide some kind of profile?' Another question to ask would be whether DDL has the same appeal for students with different cognitive styles. Reports in the literature have

tentatively suggested that field-dependent students who thrive in cooperative, interactive settings may like this inductive approach, whereas field-independent learners who are known to prefer instruction emphasising rules may not take to this kind of pedagogy. But these reports are only suggestive, at present, and need more verification. Boulton's case study has thus set the scene for explorations into a host of other corpus-based pedagogic applications to evaluate from an empirical standpoint the efficacy of DDL (see Section 7.3.3).

8.9 Evaluation of a bilingual corpus

Summary 8.9

Fan, M. and Xu, X. (2002) An evaluation of an online bilingual corpus for the self-learning of legal English. *System* 30 (1): 47–63.

This study reports the evaluation of a bilingual corpus of legal and documentary texts in English and Chinese by students doing a BA degree in Translation and Chinese. The study also sought to evaluate the usefulness of the corpus in the self-learning of English. Two comprehension tasks, a questionnaire and follow-up interview, were used for data collection.

The utility of English corpora in the field of second language learning is now gaining increasing momentum (see Section 7.2.2). Likewise, the usefulness of bilingual corpora for translator training is now widely accepted (see Section 7.5). However, an area which has not received much attention is the role of corpora for translators-in-training for purposes of improving language proficiency in specialist genres (see Bernardini et al. 2003, for more discussion on this issue). This case study addresses this topic and seeks to explore the usefulness of bilingual corpora of legal texts in the comprehension of legal language, specifically ordinances.

8.9.1 Aims

The aims of the study were threefold:

- What are the views of the students who have used the online bilingual programme?
- How do students use the online bilingual texts?
- Can the bilingual texts on ordinances solve all the comprehension problems of the students and why?

8.9.2 Corpora and methodology

The structure of the online bilingual programme included three components: an online bilingual corpus, hyperlinks and a concordancer. The online bilingual

corpus was constructed using well-matched texts in English and Chinese downloaded from the Web. The collection consisted of 100 files in each language (300,000 words in English and 500,000 characters in Chinese). The files were categorised into six subgroups: legal texts, government documents, public speeches, minutes of meetings, annual reports and press releases. Hyperlinks were inserted at the sentence level between parallel texts. A Web browser enabled the corresponding version of each individual text to be displayed on the same screen and concordancing facilities were also available.

Participants in this study were a group of 21 Year Three students doing a BA degree in Chinese and Translation at a tertiary institution in Hong Kong. Importantly, they had three courses related to law including laws for translation, translation for legal work and legal and documentary English, the focus of this case study.

The data collection consisted of a questionnaire, group/interview discussion and comprehension tasks. Two comprehension tasks were devised to create situations in which students had to use the online bilingual programme to solve legal problems; one task was based on the case of a disputed will and the other a divorce case. Both tasks required students to refer to some relevant ordinances using the programme. Students' responses were graded 'correct', 'partially correct' and 'incorrect' based on whether the students' answers showed evidence of their understanding of the ordinances aided by consulting the corpus. The aim of the questionnaire was to find out how students used the bilingual corpus when engaged in the comprehension tasks and how useful they perceived the programme. Follow-up group discussions were conducted to probe deeper into students' use and perception of the bilingual corpus.

8.9.3 Results and analysis

In the questionnaire students were asked to evaluate the three components of the online bilingual programme. Questionnaire results and follow-up interviews revealed students' views, in general, to be positive. Students appreciated the layout of the online corpus and being able to see parallel texts in separate frames on the same screen and also being able to access both versions on the Web through the hyperlinks. Students also valued the usefulness of viewing multiple examples of 'shall', and commented that the context enabled them to differentiate the meanings.

Students were also asked to comment on how they used the online bilingual texts in doing the comprehension tasks. All 21 students reported that they consulted the Chinese version first, with around 70 per cent saying that they used mainly Chinese but also the English texts when they did the exercises.

Interestingly, when it came to usefulness, 85 per cent of students considered both Chinese and English more useful, despite their reliance on Chinese. Some students commented that it was easier to see the meaning relations in the

English texts as some of the translated sentences in the Chinese version were cumbersome giving rise to comprehension problems, especially those involving postmodification of long noun phrases. The second comprehension question on ordinance of wills proved difficult with only 60 per cent of students providing correct answers. More probing follow-up questioning revealed that even when students consulted both the Chinese and English versions for definitions in the ordinances, they still could not understand the information, commenting that it was necessary to have legal knowledge to comprehend law. The authors then analysed in detail students' comprehension problems with one particular section of the ordinance on wills. A linguistic analysis of one legal sentence showed it to be very complex, consisting of a subordinate clause, a main clause including three coordinate clauses with numerous postmodifications and adverbials inserted at various points in the sentence. An examination of students' work revealed that the complicated syntactic structure of the legal sentence accounted for a significant portion of the comprehension problems reported by the students. Students' writing revealed problems at the clausal level, with a failure to spot the main clause of the sentence comprising three parallel coordinate clauses and change of subjects in these three coordinate clauses.

8.9.4 Commentary

The majority of case studies on using bilingual corpora with translation students discuss the use of corpora with regard to translation needs. This article is somewhat different as the main aim of the research was to examine the usefulness of bilingual corpora for answering comprehension questions on two legal cases. This case study has raised some interesting points related to ESP texts which have implications for ESP and subject knowledge pedagogy. Although the translation students had some background in law from their subject courses, some students could not understand the ordinances even though they had access to the Chinese corpus. In addition to comprehension problems based on lack of legal knowledge, other problems were found to be rooted in decoding of the linguistic structure of the text.

This is a small-scale study with only 21 students participating. Nevertheless, their comments and writing responses have provided some illuminating insights into their use of bilingual corpora. It also has to be borne in mind that this is a qualitative study looking at students' *evaluation* of using corpora for answering comprehension questions rather than a quantitative study on their actual writing *performance*.

8.9.5 Further research

One noteworthy aspect of the study was the authors' observation that comprehension was stymied by students' inability to process cognitively complex

linguistic structures. This case study thus suggests the need for a corpus-based action research project using concordancers for linguistic analyses of complex sentence structures. Given that students' problems were at the clause level and nominal group level involving complicated pre-and post-modification, working with a corpus tagged for different clause types and nominal groups may be beneficial. Conrad (1999) has called for corpus-based pedagogy to target more complex structures and this case study has unveiled the kind of structures that could be targeted.

However, a focus on complex language structures is just one part of the equation; competence in legal communication also requires competence in situated literacy practices, which could be given more prominence in future corpus studies. As Hafner (2008) emphasises, law graduates need discourse competence which is not solely limited to the study of generic structure potential or associated grammar and lexis, but is also situated within the practices of text construction relevant to the discourse community. In Hafner's exemplary study a discourse practitioner was 'embedded' as it were, in a legal academic community in order to engage in the kind of 'thick participation' and 'joint problematisation' envisaged by Candlin and Sarangi (2004) for the design and implementation of targeted online corpus-based materials.

While there is a dearth of studies such as the one by Fan and Xu (2002) on students' self-reports evaluating their use of corpora, there are even fewer which look at students' *performance*. One aspect of Hafner's (ibid.) study was to evaluate students' online usage patterns, relating their online behaviour to the legal writing task they were engaged in. For example, patterns of access over time and preferred online learning activities were gleaned from server logs. Such data revealed students' preference for resources and learning activities which they perceived to be relevant to their real-world goals and objectives. This goal-directed nature of students' learning was also reflected in their browsing patterns, with two different styles identified (a 'casual browse' and a 'goal-directed search'). There is a need for more research on the strategic actions students take in corpus-based learning and in this respect Hafner's thesis provides a wealth of useful ideas for further research in this area.

8.10 Creation and evaluation of a Needs-Driven Spoken Corpus for academic seminars

Summary 8.10

Jones, M. and Schmitt, N. (2010) Developing materials for discipline-specific vocabulary and phrases in academic seminars. In. N. Harwood (ed.) *English Language Teaching Materials. Theory and Practice*. Cambridge: Cambridge University Press.

In this article, corpus-based pedagogy is situated within the wider paradigm of syllabus design, moving from initial needs analysis through materials design to evaluation. The focus of the study is on discipline-specific vocabulary and phrases in academic seminars.

A number of studies have investigated the various types of lexis used in academic lectures and seminars. However, the findings from such studies tend to remain at the level of implications only; this account is noteworthy for its principled approach to the design and evaluation of corpus-inspired vocabulary materials.

8.10.1 Aims

A key aim of this study was to evaluate the effectiveness of corpus-inspired vocabulary materials used in academic seminars through posttests.

8.10.2 Corpora and methodology

The impetus for the corpus compilation came from initial needs analysis questionnaire and interview surveys among international and home students, as well as members of staff from selected departments at the University of Nottingham. The Needs-Driven Spoken Corpus (NDSC) comprises several EAP genres, the academic seminar being the focus of this article. Seminars from the School of English, the Business School and the School of Law were chosen because of the large numbers of international students they attract.

Key vocabulary in the academic seminar data was identified on the following basis. First, the seminar data were POS (part-of-speech) and semantically tagged using *WMatrix* (Rayson 2005, 2008). These data were then compared with the BNC Spoken Sampler of general English to identify categories that were over-represented in the academic seminar corpus. After statistical processing, a number of grammatical categories were found to occur significantly more often than in the BNC Spoken Sampler. Over-represented grammatical categories were found to include: single common nouns, plural common nouns, articles and adjectives. Jones and Schmitt comment that these four categories are important elements of phrases where the noun is the noun-phrase head, stating that it may be useful for students to understand these as chunks. Such nouns were therefore chosen for pedagogic treatment based on the frequency lists from the corpus analysis. However, their usefulness was also another criterion, as decided intuitively by the subject lecturers with follow-up consultation with students. Three types of vocabulary and phrases were selected for inclusion in materials for business studies: technical (e.g. *entrepreneurial audit*), general (e.g. *state of flux*), and colloquial (e.g. *stuff like that*). The vocabulary selected was then embedded into learning materials on CD-Rom, with clear explanations for use in either guided feedback sessions or independently.

8.10.3 Results and analysis

The effectiveness of the corpus-inspired vocabulary materials was evaluated through a small-scale study, proceeding through various stages. The learning-treatment consisted of two teacher-led sessions using worksheet tasks looking at types of interaction and speech acts in seminars (see Jones and Schmitt for sample worksheets). Before the second teacher-led session two weeks later, students were advised to use the programme independently to consolidate what they had learned in the first teacher-led session. An adapted C-test was administered, designed to measure whether students were able to remember the target words and phrases from the various extracts they had listened to on the CD-Rom and read in the transcripts on the worksheets.

Example 8.1 Extract of a sample C-test

> In the Language and Gender seminar, the lecturer asked the students to look through some data in the form of transcripts of conversations and identify the language features used by male and female speakers. At the beginning of the seminar, the students gave their views, using their *g____ feel_____*on the type of language used by speakers within different groups: *sin_____ se____ groups*, *mix_____ -se___ groups*, *pe_____ groups* and *ag_____ groups*.
>
> (Jones and Schmitt 2010: 244)

The results of the posttests were favourable, demonstrating considerable acquisition of the target words and phrases and thereby the effectiveness of the corpus-inspired CD-Rom programme. However, as the authors caution, the fact that students were able to produce key vocabulary and phrases in the C-test does not guarantee the same results in a much less controlled environment such as real seminars.

8.10.4 Commentary

This small-scale project stands out on several accounts. First, the vocabulary and phrases targeted for pedagogic treatment are not based solely on corpus frequency counts. Jones and Schmitt have also consulted lecturers and students for their input. In a similar vein, Ellis et al. (2008) also point out that in order to derive a pedagogically valid academic formulas list it is necessary to go beyond frequencies and complement the corpus-derived information with research tapping into their pedagogic relevance, with suggestions from teachers and students. Another aspect to consider would be to assess the psycholinguistic salience of such items, as in the study by Ellis et al. (2008). At present, as noted in

Section 7.3, there are very few reports in the literature on empirical evaluation of the effectiveness of materials so this study makes a valuable contribution to the field in this respect.

Jones and Schmitt chose to concentrate on three types of vocabulary, namely technical, general and colloquial, noting the importance of discipline-specific lexis in academic spoken discourse. There is much discussion in the literature on what constitutes technical vs. sub-technical vocabulary. The type of vocabulary classified as 'technical' by Jones and Schmitt, from their examples, e.g. *economic sanction(s), a regional body*, might equally be seen as 'semi-technical' or 'sub-technical lexis', as they are neither strictly technical nor domain-exclusive. However, the items are specialised in the sense that they are fundamental concepts in business lectures and seminars and on these grounds merit inclusion in a syllabus. But whether highly technical terms should be within the remit of the ESP teacher (even though they may be statistically significant in corpus frequencies) is open to debate with Gnutzmann (2009: 528) advocating that 'their mediation should be the prerogative of the subject teacher unless the LSP teacher also happens to be a specialist in the respective subject'.

8.10.5 Further research

The authors conclude their article by saying 'We hope this blueprint of methodology can be useful for EAP teachers to try in their own teaching situations' (p. 243) and this innovative study has opened up several avenues for further exploration. Similar studies could be conducted with a larger cohort of students (there were only 24 participants in this study). Unfortunately, the authors were not able to administer pre-tests to determine students' previous knowledge, which would be another way in which the study could be expanded. Colloquial language has hardly been touched on in corpus studies carried out in EAP/ESP environments. That this study highlights its importance in seminars and includes it in the materials on the basis of corpus frequency and also international students' lack of exposure to it is commendable. Moreover, colloquial language, e.g. *gut feeling*, *blokey*, to cite instances from the authors' list, very often involves the use of metaphor, which would be another interesting aspect for further research.

8.11 Conclusion

This section has showcased some studies which are representative of the wide range of corpus research undertaken to date and its applications. The selection of cases also represents the myriad perspectives from which the lexis, grammar and lexico-grammar can be analysed: critical discourse-analytical, rhetorical/genre-based, intercultural, sociolinguistic, pragmatic and ethnographic. The

cases also exemplify the different software available, e.g. *WordSmith Tools*, *WMatrix*, *SketchEngine*, *ConcGram*, and the respective functions of these tools. Through this exposition of ten quite different corpus studies, my aim has been to suggest how novice researchers might develop and adapt the choice of corpora and methods of these cases for their own research projects. It is hoped that the 'Further research' sections will provide some impetus and inspiration for more innovative work in this burgeoning field.

Part IV
Resources

9
Key Sources

The following sections cover a wide range of valuable resources which, I hope, will provide readers with a good starting point from which to explore the wealth of information available on corpus linguistics.

9.1 Books

The main references in specific areas of corpus linguistics have been listed at the ends of relevant chapters in this book, together with a brief description of each one. Those which are key general sources are given again below. The volumes by Biber et al. (1998), McEnery et al. (2006), Meyer (2002) and O'Keeffe et al. (2007) also contain appendices listing information on individual corpora. The guide by Baker et al. (2006) provides an extensive list of important corpora, together with their common acronyms.

Baker, P., Hardie, A. and McEnery, T. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics. Investigating Language Structure and Language Use*. Cambridge: Cambridge University Press.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics*, 2nd edn. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies. An Advanced Resource Book*. London: Routledge.

Meyer, C. (2002) *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Teubert, W. and Čermáková, A. (2004). *Corpus Linguistics. A Short Introduction*. London: Continuum.

9.2 Edited collections

Aarts, J., de Haan, P. and Oostdijk, N. (eds) (1993) *English Language Corpora*. Amsterdam: Rodopi.
Ädel, A. and Reppen, R. (eds) (2008) *Corpora and Discourse: the Challenges of Different Settings*. Amsterdam: John Benjamins.
Aijmer, K. (ed.) (2009) *Corpora and Language Teaching*. Amsterdam: John Benjamins.
Aijmer, K. and Altenberg, B. (eds) (1991) *English Corpus Linguistics: Studies in Honor of Jan Svartvik*. London: Longman.
Aijmer, K. and Altenberg, B. (eds) (2004) *Advances in Corpus Linguistics*. Amsterdam: Rodopi.
Aijmer, K., Altenberg, B. and Johansson, M. (eds) (1996) *Languages in Contrast*. Lund: Lund University Press.
Aijmer, K. and Stenström, A.-B. (eds) (2004) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins.
Altenberg, B. and Granger, S. (eds) (2002) *Lexis in Contrast: Corpus-Based Approaches*. Amsterdam: John Benjamins.
Aston, G., Bernardini, S. and Stewart, D. (eds) (2004) *Corpora and Language Learners*. Amsterdam: John Benjamins.
Baker, M., Francis G. and Tognini-Bonelli, E. (eds) (1993) *Text and Technology*. Amsterdam: John Benjamins.
Baker, P. (ed.) (2009) *Contemporary Corpus Linguistics*. London: Continuum.
Beeby, S., Rodríguez Inés, P. and Sánchez-Gijón, P. (eds) (2009) *Corpus Use and Translating*. Amsterdam: John Benjamins.
Biber, D., Connor, U. and Upton, T. (eds) (2007) *Discourse on the Move*. Amsterdam: John Benjamins.
Bondi, M. and Scott, M. (eds) (2010) *Keyness in Texts*. Amsterdam: John Benjamins.
Braun, S., Kohn, K. and Mukherjee, J. (eds) (2006) *Corpus Technology and Language Pedagogy*. Frankfurt am Main: Peter Lang.
Burnard, L. and McEnery, T. (eds) (2000) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang.
Campoy-Cubillo, M., Bellés-Fortuño, B. and Gea-Valor, M. (eds) (2010) *Corpus-Based Approaches to English Language Teaching*. London: Continuum.
Campoy-Cubillo, M. and Luzón, M. J. (eds) (2007) *Spoken Corpora in Applied Linguistics*. Berlin: Peter Lang.
Charles, M., Pecorari, D. and Hunston, S. (eds) (2009) *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum.

Coffin, C., Hewings, A. and O'Halloran, K. (eds) (2004) *Applying English Grammar: Functional and Corpus Approaches*. The Open University: Arnold.

Connor, U. and Upton, T. (eds) (2002) *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. Amsterdam: Rodopi.

Connor, U. and Upton, T. (eds) (2004a) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins.

Connor, U. and Upton, T. (eds) (2004b) *Applied Corpus Linguistics: a Multidimensional Perspective*. Amsterdam: Rodopi.

Frankenberg-Garcia, A., Flowerdew, L. and Aston, G. (eds) (2011) *New Trends in Corpora and Language Learning*. London: Continuum.

Ghadessy, M. Henry, A. and Roseberry, R. (eds) (2001) *Small Corpus Studies and ELT*. Amsterdam: John Benjamins.

Gilquin, G., Papp, S. and Díez-Bedmar, M. (eds) (2008) *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi.

Granger, S. (ed.) (1998) *Learner English on Computer*. London: Longman.

Granger, S., Hung, J. and Petch-Tyson, S. (eds) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S., Lerot, J. and Petch-Tyson, S. (eds) (2003) *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Granger, S. and Meunier, F. (eds) (2008) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

Gries, S. Th. and Stefanowitsch, A. (eds) (2006) *Corpora in Cognitive Linguistics*. Berlin: Mouton de Gruyter.

Hidalgo, E., Quereda, L. and Santana, J. (eds) (2007) *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.

Hoey, M. (ed.) (1993) *Data, Description, Discourse*. London: HarperCollins.

Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W. (eds) (2007) *Text, Discourse and Corpora*. London: Continuum.

Hornero, A., Luzón, L. and Murillo, S. (eds) (2006) *Corpus Linguistics: Applications for the Study of English*, pp. 301–12. Bern: Peter Lang.

Hundt, M., Nesselhauf, N. and Biewer, C. (eds) (2007) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Hyland, K., Chau, M.H. and Handford, M. (eds) (in press, 2012) *Corpora in Applied Linguistics: Current Approaches and Future Directions*. London: Continuum.

Jucker, A., Schreier, D. and Hundt, M. (eds) (2009) *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi.

Kawaguchi, Y., Minegishi, M. and Durand, J. (eds) (2009) *Corpus Analysis and Variation in Linguistics*. Amsterdam: John Benjamins.

Kettemann, B. and Marko, G. (eds) (2002) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi.

Leistyna, P. and Meyer, C. (eds) (2003) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi.
Ljung, M. (ed.) (1997) *Corpus-Based Studies in English*. Amsterdam: Rodopi.
Mair, C. and Hundt, M. (eds) (2000) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.
Melia, J. and Lewandowska, B. (eds) (1997) *Practical Applications in Language Corpora*. University of Lodz, Poland.
Meunier, F. and Granger, S. (eds) (2008) *Phraseology in Foreign Language Teaching and Learning*. Amsterdam: John Benjamins.
Morley, J. and Bailey, P. (eds) (2009) *Corpus-Assisted Discourse Studies on the Iraq Conflict*. London: Routledge.
Parodi, G. (ed.) (2007) *Working with Spanish Corpora*. London: Continuum.
Parodi, G. (ed.) (2010) *Academic and Professional Discourse Genres in Spanish*. Amsterdam: John Benjamins.
Partington, A., Morley, J. and Haarman, L. (eds) (2004) *Corpora and Discourse*. Bern: Peter Lang.
Percy, C., Meyer, C. and Lancashire, I. (eds) (1996) *Synchronic Corpus Linguistics*. Amsterdam: Rodopi.
Peters, P., Collins, P. and Smith, A. (eds) (2002) *New Frontiers of Corpus Research*. Amsterdam: Rodopi.
Renouf, A. and Kehoe, A. (eds) (2006) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.
Renouf, A. and Kehoe, A. (eds) (2009) *Corpus Linguistics. Refinements and Reassessments*. Amsterdam: Rodopi.
Reppen, R., Fitzpatrick, S. and Biber, D. (eds) (2002) *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.
Römer, U. and Schulze, R. (eds) (2009) *Exploring the Lexis–Grammar Interface*. Amsterdam: John Benjamins.
Romero-Trillo, J. (ed.) (2008) *Pragmatics and Corpus Linguistics*. New York: Mouton de Gruyter.
Saito, T., Nakamura, J. and Yamazaki, S. (eds) (2002) *English Corpus Linguistics in Japan*. Amsterdam: Rodopi.
Sampson, G. and McCarthy, D. (eds) (2004) *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.
Schneider, K. and Barron, A. (eds) (2008) *Variational Pragmatics*. Amsterdam: John Benjamins.
Scott, M. and Thompson, G. (eds) (2001) *Patterns of Text*. Amsterdam: John Benjamins.
Simpson, R. and Swales, J. (eds) (2001) *Corpus Linguistics in North America*. Ann Arbor, Mich.: University of Michigan Press.
Sinclair, J. McH. (ed.) (1987) *Looking Up. An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. McH. (ed.) (2004) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
Sinclair, J. McH., Hoey, M. and Fox, G. (eds) (1993) *Techniques of Description – Spoken and Written Discourse*. London: Routledge.
Svartvik, J. (ed.) (1992) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.
Thomas, J. and Boulton, A. (eds) (2012) *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.
Thomas, J. and Short, M. (eds) (1996) *Using Corpora for Language Research*. London: Longman.
Thompson, G. and Hunston, S. (eds) (2006) *System and Corpus: Exploring Connections*. Equinox.
Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds) (1997) *Teaching and Language Corpora*. London: Longman.
Wilson, A., Archer, D. and Rayson, P. (eds) (2006) *Corpus Linguistics around the World*. Amsterdam: Rodopi.
Zanettin, F., Bernardini, S. and Stewart, D. (eds) (2003) *Corpora in Translator Education*. Manchester, UK: St Jerome Publishing.

9.3 Handbooks

Handbooks dedicated to corpus linguistics are as follows:

Lüdeling, A. and Kytö, M. (eds) (2008) *Corpus Linguistics. An International Handbook*. Volumes 1 and 2. Berlin: Walter de Gruyter.
O'Keeffe, A. and McCarthy, M. (eds) (2010) *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
Teubert, W. and Krishnamurthy, R. (eds) (2007) *Corpus Linguistics* (6 volumes). Critical Concepts in Linguistics Series. London: Routledge.

The following handbooks contain chapters on various aspects of corpus linguistics:

Aarts, B. and McMahon, A. (eds) (2006) *The Handbook of English Linguistics*. Oxford, UK: Wiley-Blackwell.
Davies, A. and Elder, C. (eds) (2005) *The Handbook of Applied Linguistics*. Malden: Mass.: Blackwell.
Gee, J.P. and Handford, M. (eds) (2011) *The Routledge Handbook of Discourse Analysis*. London: Routledge.

9.4 Journals

9.4.1 Corpus linguistic journals

- *Corpora*
  A wide variety of articles ranging from research findings based on the exploitation of corpora to accounts of corpus building, corpus tool construction

and corpus annotation schemes can be found in this journal. The journal is interdisciplinary in scope and also features articles on other languages besides English.
http://www.euppublishing.com/journal/cor

- *Corpus Linguistics and Linguistic Theory*
  The articles in this journal are of a theoretical nature and address issues in all core areas of linguistic research, e.g. phonology, morphology, syntax, semantics, pragmatics and cognitive linguistics.
  http://www.reference-global.com/loi/cllt
- *ICAME Journal* (*International Computer Archive of Modern and Medieval English*)
  ICAME mostly publishes descriptive studies on linguistic research based on modern and historical corpora.
  http://icame.uib.no/journal.html
- *International Journal of Corpus Linguistics*
  Articles in this journal focus on empirical research covering applied and theoretical work in all areas of corpus linguistics, e.g. lexicology, grammar, discourse analysis. The journal also interfaces with computational linguistics and publishes articles which address corpus methodologies.
  http://www.benjamins.nl/cgi-bin/t_seriesview.cgi?series=ijcl
- *Literary and Linguistic Computing*
  This journal publishes articles on all aspects of computing and information technology applied to literature and language research and teaching.
  http://www.oxfordjournals.org/our_journals/litlin/about.html

9.4.2 Related journals

- *Applied Linguistics*
  http://applij.oxfordjournals.org/content/by/year
- *English for Specific Purposes*
  http://elsevier.com/locate/esp
- *English Text Construction*
  http://www.ingentaconnect.com/content/jbp/etc
- *Journal of Applied Linguistics and Professional Practice*
  http://www.equinoxpub.com/JALPP
- *Journal of English for Academic Purposes*
  http://elsevier.com/locate/jeap
- *Journal of Pragmatics*
  http://elsevier.com/locate/pragma
- *Language Learning and Technology*
  http://llt.msu.edu/
- *ReCall*
  http://journals.cambridge.org/action/displayJournal?jid=REC

- *System*
  http://elsevier.com/locate/system/
- *TESOL Quarterly*
  http://tesol.org/s_tesol/seccss.asp?CID=632&DID=2461
- *Text and Talk*
  http://degruyter.de/journals/text/detailEn.cfm

9.5 Principal corpus linguistic conferences and associations

- *AACL* (*American Association for Corpus Linguistics*) *Conference*
  http://www2.gsu.edu/~wwwesl/alesl/aacl2011
- *ICAME* (*International Computer Archive of Modern and Medieval English) Conference*
  http://icame.uib.no/
- *International Conference on Corpus Linguistics* (Spanish Association for Corpus Linguistics)
  http://www.cilc2011.upv.es
- *International Corpus Linguistics Conference*
  http://cl2011.org.uk/archives.html
- *IVACS* (*The Inter-Varietal Applied Corpus Studies*) *Conference*
  http://www.mic.ul.ie/ivacs/symposium.htm
- *Practical Applications in Language and Computers Conference*
  http://palc.ia.uni.lodz.pl/
- *TaLC* (*Teaching and Language Corpora*) *Conference*
  http://talc8.isla.pt/
- *TELRI* (*Trans-European Language Resources Infrastructure*) *Seminars*
  http://telri.nytud.hu/telri2/seminar/seminar.html

9.5.1 SIGs (special interest groups)

- ACL (Association for Computational Linguistics) SIGWAC, Special Interest Group on Web as Corpus
  http://www.sigwac.org.uk/
- BAAL (British Association for Applied Linguistics) Corpus Linguistics SIG
  http://corpus-sig-baal.org.uk/
- EUROCALL (European Association for Computer-Assisted Language Learning) CorpusCALL SIG
  http://www.corpuscall.org.uk/
- IATEFL (International Association of Teachers of English as a Foreign Language) Business English SIG covers the use of corpora in teaching business English
  http://www.iatefl.org/special-interest-groups/sigs
- PALA (Poetics and Linguistics Association) Corpus Stylistics SIG
  http://www.pala.ac.uk/resources/sigs/corpus-style/

9.6 Key Internet sites

- David Lee's 'Bookmarks for Corpus-based Linguists' (http://tiny.cc/corpora) is a useful starting place for accessing information on the following: corpora, data archives, non-English, parallel and multilingual corpora, software and tools, references and articles, corpus-based language teaching, conferences and projects.
- http://ucrel.lancs.ac.uk. The homepage of the University Centre for Computer Research on Language (UCREL) at the University of Lancaster, UK provides links to mailing lists, book series and journals, and research groups/projects in corpus linguistics and natural language processing in the UK, Europe and other parts of the world.
- http://www.uclouvain.be/en-cecl-icle.html. This site gives information on the learner corpus projects being carried out at the University of Louvain. It also provides a learner corpus bibliography and links to learner corpus projects carried out worldwide.
- http://www.helsinki.fi/varieng/CoRD/. The Corpus Resource Database (CoRD), maintained by the Research Unit for Variation, Contacts and Change in English at the University of Helsinki, is an open access online resource through which academic corpus compilers can make available basic information about their corpora.
- http://ota.ahds.ac.uk/. The Oxford Text Archive (OTA) develops, collects, catalogues and preserves electronic literary and linguistic resources for use in higher education, in research, teaching and learning.
- http://www.athel.com/corpus.html. This site developed by Michael Barlow has information on learner corpora, parallel corpora, corpus-based language learning, corpus software and links to other useful sites.

9.7 Sites for concordancers, search engines and text-analysis tools

AntConc http://antlab.sci.waseda.ac.jp/
BNCWeb http://www.bncweb.info/
Compleat Lexical Tutor http://www.lextutor.ca/concordancers/
ConcGram http://repository.lib.polyu.edu.hk/jspui/handle/10397/601
KfNgram www.kwicfinder.com/KfNgram
MonoConc http://www.athel.com/corpus.html
SketchEngine http://www.sketchengine.co.uk
UAM Corpus Tools http://wagsoft.com/CorpusTool/UAM
WebCorp http://www.webcorp.org.uk/

WordSmith Tools http://www.lexically.net/wordsmith/index.html
WMatrix http://ucrel.lancs.ac.uk/wmatrix/

Appendices in the following books contain overviews of a range of software tools for both research and teaching purposes.

Adolphs, S. (2006) *Introducing Electronic Text Analysis*. London: Routledge.
Bennett, G. (2010) *Using Corpora in the Language Learning Classroom*. Ann Arbor, Mich.: University of Michigan Press.
Bowker, L. and Pearson, J. (2002) *Working with Specialised Language. A Practical Guide to Using Corpora*. London: Routledge.
Reppen, R. (2010) *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

9.8 E-mail lists

- Corpora-List is the main discussion list for corpus linguistics. It also has postings by researchers in natural language processing.
  corpora@lists.uib.no
- CLLT (Corpus Linguistics and Language Teaching) focuses on the application of corpora in language teaching.
  cllt@mailman.rice.edu
- Humanist List provides a discussion forum on humanities computing and digital humanities. It is affiliated with the Association for Computers and the Humanities (ACH) and the Association for Literary and Linguistic Computing (ALLC).
  www.digitalhumanities.org/humanist/
- Linguist List often has postings on corpus linguistic-related books and issues.
  linguist@LINGUISTLIST.ORG
- Sysfling is a discussion list devoted to systemic-functional linguistics which sometimes has contributions from corpus linguists working within an SFL framework.
  sysfling@mailman.cf.ac.uk

References

Aarts, B. (2000) Corpus linguistics, Chomsky and fuzzy tree fragments. In C. Mair and M. Hundt (eds), *Corpus Linguistics and Linguistic Theory*, pp. 5–13. Amsterdam: Rodopi.

Aarts, J. (2002a) Does corpus linguistics exist? Some old and new issues. In L. Breivik and A. Hasselgren (eds), *From the Colt's Mouth and Others*, pp. 1–17. Amsterdam: Rodopi.

Aarts, J. (2002b) Review of *Corpus Linguistics at Work*. *International Journal of Corpus Linguistics*, 7 (1): 118–23.

Aarts, J. (2006) Corpus linguistics, grammar and theory: report on a panel discussion at the 24th ICAME Conference. In A. Renouf and A. Kehoe (eds), *The Changing Face of Corpus Linguistics*, pp. 391–408. Amsterdam: Rodopi.

Ackerley, K. and Coccetta, F. (2007) Enriching language learning through a multimedia corpus. *ReCALL*, 19 (3): 351–70.

Ädel, A. (2006) *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Ädel, A. and Reppen, R. (eds), (2008) *Corpora and Discourse: the Challenges of Different Settings*. Amsterdam: John Benjamins.

Adolphs, S. (2008) *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins.

Adolphs, S., Atkins, S. and Harvey, K. (2007) Caught between professional requirements and interpersonal needs: vague language in healthcare contexts. In J. Cutting (ed.), *Vague Language Explored*, pp. 62–78. Basingstoke: Palgrave Macmillan.

Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004) Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1 (1): 9–28.

Adolphs, S. and Carter, R. (2007) Beyond the word: new challenges in analysing corpora of spoken English. *European Journal of English Studies*, 11 (2): 133–46.

Aijmer, K. (1984) *Sort of* and *kind of* in English conversation. *Studia Linguistica*, 38: 118–28.

Aijmer, K. (2002) *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.

Aijmer, K. (ed.) (2009) *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Aijmer, K. and Altenberg, B. (eds) (1991) *English Corpus Linguistics*. London: Longman.

Aijmer, K. and Altenberg, B. (eds) (2004) *Advances in Corpus Linguistics*. Amsterdam: Rodopi.

Aijmer, K. and Stenström, A.-B. (2004a) Discourse patterns in spoken and written corpora. In K. Aijmer and A.-B. Stenström (eds), *Discourse Patterns in Spoken and Written Corpora*, pp. 1–13. Amsterdam: John Benjamins.

Aijmer, K. and Stenström, A.-B. (eds) (2004b) *Discourse Patterns in Spoken and Written Corpora*. Amsterdam: John Benjamins.

Alderson, C. (1996) Do corpora have a role in language assessment? In J. Thomas and M. Short (eds), *Using Corpora for Language Research*, pp. 248–59. London: Longman.

Alderson, C. (2007) Judging the frequency of English words. *Applied Linguistics*, 28 (3): 383–409.

Alexopoulou, T. (2008) Building new corpora for English profile. *Research Notes*, 33: 15–18. Accessed 3 July, http://www.cambridgeesol.org/rs_notes/.

Ali Mohamed, A. (2007) Semantic fields of problem in business English: Malaysian and British journalistic business texts. *Corpora*, 2 (2): 211–39.

Alonso Belmonte, I. (2009) Towards a genre-based characterization of the problem–solution textual pattern in English newspaper editorials and op-eds. *Text and Talk*, 29 (4): 393–414.

Altenberg, B. (1998) On the phraseology of spoken English: the evidence of recurrent word combinations. In A. P. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford Studies in Lexicography and Lexicology, pp. 101–22. Oxford: Clarendon Press.

Altenberg, B. and Granger, S. (eds) (2002) *Lexis in Contrast: Corpus-Based Approaches*. Amsterdam: John Benjamins.

Amador Moreno, C., Chambers, A. and O'Riordan, S. (2006) Integrating a classroom of corpus discourse in language teacher education: the case of discourse markers. *ReCALL*, 18 (1): 83–104.

Andersen, G. (2010) How to use corpus linguistics in sociolinguistics. In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 547–62. London: Routledge.

Anderson, W. and Corbett, J. (2009) *Exploring English with Online Corpora. An Introduction*. Basingstoke: Palgrave Macmillan.

Anderson, W. and Corbett, J. (2010) Teaching English as a friendly language: lessons from the SCOTS Corpus. *ELT Journal*, 64 (4): 414–23.

Andor, J. (2004) The master and his performance: an interview with Noam Chomsky. *Intercultural Pragmatics*, 1 (1): 93–111.

Archer, D. (2007) Computer-assisted literary stylistics: the state of the field. In M. Lambrou and P. Stockwell (eds), *Contemporary Stylistics*, pp. 244–56. London: Continuum.

Arts and Humanities Research Council: Forensic Linguistics. Available at: http://www.ahrc.ac.uk/About/Policy/Pages/Evaluation.aspx

Aston, G. (1995) Corpora in language pedagogy. Matching theory and practice. In G. Cook and B. Seidlhofer (eds), *Principle and Practice in Applied Linguistics*, pp. 257–70. Oxford: Oxford University Press.

Aston, G. (2000) I corpora come risorse per la traduzione e per l'apprendimento. In S. Bernardini and F. Zanettin (eds), *I Corpora Nella Didattica della Traduzione*, pp. 21–30. Bologna: CLUEB.

Aston, G. and Burnard, L. (1998) *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Aston, G., Bernardini, S. and Stewart, D. (eds) (2004) *Corpora and Language Learners*. Amsterdam: John Benjamins.

Atkins, S., Clear, J. and Ostler, N. (1992) Corpus design criteria. *Literary and Linguistic Computing*, 7 (1): 1–16.

Atkins, S. and Harvey, K. (2010) How to use corpus linguistics in the study of health communication. In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 605–19. London: Routledge.

Atkins, S., Levin, B. and Zampolli, A. (1994) Computational approaches to the lexicon: an overview. In S. Atkins and A. Zampolli (eds), *Computational Approaches to the Lexicon*, pp. 17–45. Oxford: Oxford University Press.

Atkins, S. and Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Atkins, S. and Zampolli, A. (eds) (1994) *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Baker, C., Fillmore, C. and Cronin, B. (2003) The structure of the FrameNet database. *International Journal of Lexicography*, 16 (3): 281–96.

Baker, M. (1993) Corpus linguistics and translation studies: implications and applications. In M. Baker et al. (eds), *Text and Technology*, pp. 233–50. Amsterdam: John Benjamins.

Baker, M. (1995) Corpora in translation studies: an overview and some suggestions for future research. *Target*, 7 (2): 223–43.

Baker, M. (1996) Corpus-based translation studies: the challenges that lie ahead. In H. Somers (ed.), *Technology, LSP and Translation: Studies in Language Engineering*, pp. 175–86. Amsterdam: John Benjamins.

Baker, M. (2004) A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9 (2): 167–93.

Baker, M., Francis G. and Tognini Bonelli, E. (eds) (1993) *Text and Technology*. Amsterdam: John Benjamins.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. (ed.) (2009) *Contemporary Corpus Linguistics*. London: Continuum.

Baker, P. (2010) *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P., Gabrielatos, C., KhosraviNik, M., KrzyŽanowski, 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.

Baker, P., Hardie, A. and McEnery, T. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P. and McEnery, T. (2005) A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, 4 (2): 197–226.

Bakhtin, M. (1986) *Speech Genres and Other Late Essays*. Austin, Tex.: University of Texas Press.

Baldry, A. and O'Halloran, K. (in press, 2012) *Multi-Modal Corpus-Based Approaches to Website Analysis*. London: Equinox.

Baldry, A. and Thibault, P. (2001) Towards multimodal corpora. In G. Aston and L. Burnard (eds), *Corpora in the Description and Teaching of English*, pp. 87–102. Bologna: CLUEB.

Baldry, A. and Thibault, P. (2006) Multimodal corpus linguistics. In G. Thompson and S. Hunston (eds), *System and Corpus: Exploring Connections*, pp. 165–183. Equinox.

Bargiela-Chiappini, F. and Harris, S. (1997) *Managing Language: the Discourse of Corporate Meetings*. Amsterdam: John Benjamins.

Bargiela-Chiappini, F., Nickerson, C., and Planken, B. (2007) *Business Discourse*. Basingstoke: Palgrave Macmillan.

Barker, F. (2008a) Using corpora for language assessment: trends and prospects. Paper presented at The Fourth Inter-Varietal Applied Corpus Studies International Conference. University of Limerick, 13 June.

Barker, F. (2008b) Exploring lexical differences in general English reading papers. *Research Notes*, 31: 22–6. Accessed 3 July, http://www.cambridgeesol.org/rs_notes/

Barker, F. (2010) How can corpora be used in language testing? In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 633–46. London: Routledge.

Barlow, M. (1996) Corpora for theory and practice. *International Journal of Corpus Linguistics*, 1 (1): 1–37.

Barlow, M. (2000) Parallel texts in language teaching. In S. Botley, T. McEnery and A. Wilson (eds), *Multilingual Corpora in Teaching and Research*, pp. 106–15. Amsterdam: Rodopi.

Barlow, M. (2005) Computer based analyses of learner language. In R. Ellis and G. Barkhuizen, *Analysing Learner Language*, pp. 335–58. Oxford: Oxford University Press.

Barlow, M. and Kemmer, S. (eds) (2000) *Usage-Based Models of Language*. Stanford: CLSI Publications.

Barnbrook, G. (1996) *Language and Computers*. Edinburgh: Edinburgh University Press.

Bazerman, C. (1988) *Shaping Written Knowledge: the Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.

Bazerman, C. (1994) Systems of genres and the enactments of social intentions. In A. Freedman and P. Medway (eds), *Genre and the New Rhetoric*, pp. 79–101. London: Taylor and Francis.

Becker, J. (1975) The phrasal lexicon. In R. Shank and B.L Nash-Webber (eds), *Theoretical Issues in Natural Language Processing*, pp. 60–3. Cambridge, Mass.: Bolt, Beranek and Newman.

Bednarek, M. (2006) *Evaluation in Media Discourse*. London: Continuum.

Bednarek, M. (2007) Teaching English literature and linguistics using corpus stylistic methods. *Bridging Discourses: ASFLA 2007 Online Proceedings*.

Beeby, S., Rodríguez Inés, P. and Sánchez-Gijón, P. (eds) (2009) *Corpus Use and Translating*. Amsterdam: John Benjamins.

Beeching, K. (2006) Synchronic and diachronic variation: the how and why of sociolinguistic corpora. In A. Wilson, D. Archer and P. Rayson (eds), *Corpus Linguistics around the World*, pp. 49–61. Amsterdam: Rodopi.

Béjoint, H. (2010) *The Lexicography of English*. Oxford: Oxford University Press.

Belcher, D. and Hirvela, A. (eds) (2008) *The Oral–Literate Connection. Perspectives on L2 Speaking, Writing, and Other Media Interactions*. Ann Arbor, Mich.: University of Michigan Press.

Belcher, D., Johns, A. and Paltridge, B. (eds) (2011) *New Directions in English for Specific Purposes Research*. Ann Arbor, Mich.: University of Michigan Press.

Belz, J. (2004) Learner corpus analysis and the development of foreign language proficiency. *System*, 32 (4): 577–91.

Beneke, J. (1991) English as lingua franca or as medium of intercultural communication. In R. Grebing (ed.), *Grenzenloses Sprachenlernen*, pp. 54–66. Berlin: Cornelsen.

Bennett, G. (2010) *Using Corpora in the Language Learning Classroom. Corpus Linguistics for Teachers*. Ann Arbor, Mich.: University of Michigan Press.

Bernardini, S. (2000) *Competence, Capacity, Corpora. A Study in Corpus-Aided Language Learning*. Bologna: CLUEB.

Bernardini, S. (2002) Serendipity expanded: exploring new directions for discovery learning. In B. Kettermann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 165–82. Amsterdam: Rodopi.

Bernardini, S. (2004) Corpora in the classroom: an overview and some reflections on future developments. In J. Sinclair (ed.), *How to Use Corpora in Language Teaching*, pp. 15–36. Amsterdam: John Benjamins.

Bernardini, S., Stewart, D. and Zanettin, F. (2003) Corpora in translator education: an introduction. In F. Zanettin et al. (eds), *Corpora in Translator Education*, pp. 1–13. Manchester, UK: St Jerome Publishing.

Bhatia, V.K. (2008) Genre analysis, ESP and professional practice. *English for Specific Purposes*, 27 (3): 161–74.

Bhatia, V.K., Flowerdew, J. and Jones, R. (eds) (2008) *Advances in Discourse Studies*. London: Routledge.

Bhatia, V.K., Langton, N. and Lung, J. (2004) Legal discourse: opportunities and threats for corpus linguistics. In U. Connor and T. Upton (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 203–31. Amsterdam: John Benjamins.

Bianchi, F. and Pazzaglia, R. (2007) Student writing of research articles in a foreign language: metacognition and corpora. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years On*, pp. 259–87. Amsterdam: Rodopi.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1990) Methodological issues regarding corpus-based analysis of linguistic variation. *Literary and Linguistic Computing*, 5 (4): 257–69.

Biber, D. (2003) Variation among university spoken and written registers: a new multidimensional analysis. In P. Leistyna and C. Meyer (eds), *Corpus Analysis: Language Structure and Language Use*, pp. 47–67. Amsterdam: Rodopi.

Biber, D. (2004) A corpus linguistic investigation of vocabulary-based discourse units in university registers. In U. Connor and T. Upton (eds), *Applied Corpus Linguistics: a Multidimensional Perspective*, pp. 53–72. Amsterdam: Rodopi.

Biber, D. (2006) *University Language. A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D. (2009) A corpus-driven approach to formulaic language in English: multiword patterns in speech and writing. *International Journal of Corpus Linguistics*, 14 (3): 275–311.

Biber, D., Connor, U. and Upton, T. (eds) (2007) *Discourse on the Move*. Amsterdam: John Benjamins.

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., Conrad, S., and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P. and Helt, M. (2002) Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly*, 36 (1): 9–48.

Biber, D., Conrad, S., Reppen, R., Byrd, P., and Helt, M. (2003) The authors respond: strengths and goals of multidimensional analysis. *TESOL Quarterly*, 37 (1): 151–5.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V. and Urzua, A. (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Corpus*. TOEFL Monograph Series. ETS.

Biber, D. and Jamieson, J. (1998) *Final Report: Pilot Study to Test the Influence of Linguistic Variables on Listening and Reading Test Performance*. Princeton, NJ: Educational Testing Service.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Bjørge, A. K. (2010) Conflict or cooperation: the use of backchanneling in EFL negotiations. *English for Specific Purposes*, 29 (3): 191–203.

Bley-Vroman, R. (1983) The comparative fallacy in interlanguage studies. The case of systematicity. *Language Learning*, 33: 1–17.

Bloch, J. (2009) The design of an online concordancing program for teaching about reporting verbs. *Language Learning and Technology*, 13 (1): 59–78.

Blommaert, J. (2005) *Discourse*. Cambridge: Cambridge University Press.

Bloor, M. and Bloor, T. (1986) Language for specific purposes: practice and theory. In CLCS Occasional Papers. Dublin: Centre for Language and Communication Studies, Trinity College.

Bolt, P. and Bolton, K. (1996) The international corpus of English in Hong Kong. In S. Greenbaum (ed.), *Comparing English Worldwide: the International Corpus of English*, pp. 197–214. Oxford: Clarendon Press.

Bolton, K. (2003) *Chinese Englishes: a Sociolinguistic History*. Cambridge: Cambridge University Press.

Bolton, K. and Kachru, B. (eds) (2006) *World Englishes; Critical Concepts in Linguistics*. London: Routledge.

Bondi, M. and Scott, M. (eds) (2010) *Keyness in Texts*. Amsterdam: John Benjamins.

Bosseux, C. (2004) Point of view in translation: a corpus-based study of French translations of Virginia Woolf's *To the Lighthouse*. *Across Languages and Cultures*, 5 (1): 107–22.

Botley, S., McEnery, T. and Wilson, A. (eds) (2000) *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.

Boulton, A. (2007) But where's the proof? The need for empirical evidence for data-driven learning. *Proceedings of the BAAL Conference 2007*.

Boulton, A. (2009) Testing the limits of data-driven learning: language proficiency and training. *ReCALL*, 21 (1): 37–54.

Boulton, A. (2010) Data-driven learning: taking the computer out of the equation. *Language Learning*, 60 (3): 534–72.

Bowker, L. and Pearson, J. (2002) *Working with Specialized Language: a Practical Guide to Using Corpora*. London: Routledge.

Bowles, H. (2006) Bridging the gap between conversation analysis and ESP – an applied study of the opening sequences of NS and NNS service telephone calls. *English for Specific Purposes*, 25 (3): 332–57.

Brand, C. and Götz, S. (2011) Fluency vs. accuracy in advanced spoken learner language: a multi-method approach. *International Journal of Corpus Linguistics* 16 (2): 255–75.

Braun, S. (2006) ELISA: a pedagogically enriched corpus for language learning purposes. In S. Braun et al. (eds), *Corpus Technology and Language Pedagogy*, pp. 25–47. Frankfurt am Main: Peter Lang.

Braun, S. (2007) Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora. *ReCALL*, 19 (3): 307–28.

Braun, S., Kohn, K. and Mukherjee, J. (eds) (2006) *Corpus Technology and Language Pedagogy*, pp. 25–47. Frankfurt am Main: Peter Lang.

Brazil, D. (1995) *A Grammar of Speech*. Oxford: Oxford University Press.

Bréal, M. (1897) *Essai de sémantique*. Paris: Hachette.

Breyer, Y. (2006a) *My concordancer*: tailor-made software for language learners and teachers. In S. Braun et al. (eds), *Corpus Technology and Language Pedagogy*, pp. 157–76. Frankfurt am Main: Peter Lang.

Breyer, Y. (2006b) How to teach with corpora: integrating corpora into initial teacher training. Paper given at the 7th Teaching and Language Corpora Conference. Paris, France, 1–4 July.

Brook, G.L. (1970) *The Language of Dickens*. London: André Deutsch.

Brown, P. and Levinson, S. (1987) *Politeness*. Cambridge: Cambridge University Press.

Burgess, G. (2000) Corpus analysis in the service of literary criticism: Goethe's *Die Wahlverwandtschaften* as a model case. In B. Dodd (ed.), *Working with German Corpora*, pp. 40–68. Birmingham: The University of Birmingham Press.

Burnard, L. and McEnery, T. (eds) (2000) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang.

Butler, C.S. (1998) Multi-word lexical phenomena in functional grammar. *Revista Canaria de Estudios Ingleses*, 36: 13–36.

Butler, C.S. (2003a) Multi-word sequences and their relevance for recent models of functional grammar. *Functions of Language*, 10 (2): 179–208.

Butler, C.S. (2003b) Review of *Corpus Linguistics at Work*. *System*, 31 (1): 128–32.

Butler, C.S. (2004a) Formulaic language: corpus-based and psycholinguistic approaches. Workshop given at the 6th TALC Conference, Granada, 6 July 2004.

Butler, C.S. (2004b) Corpus studies and functional linguistic theories. *Functions of Language*, 11 (2): 147–86.

Butler, S. (1997) Corpus of English in Southeast Asia: implications for a regional dictionary. In M.L.S. Bautista (ed.), *English is an Asian Language*: *the Philippine Context*, pp. 103–24. Manila: The Macquarie Library.

Caldas-Coulthard, C. (1993) From discourse analysis to critical discourse analysis: the differential re-presentation of women and men speaking in written news. In J. McH. Sinclair et al. (eds), *Techniques of Description – Spoken and Written Discourse*, pp. 196–208. London: Routledge.

Callies, M. (2009) *Information Highlighting in Advanced Learner English*. Amsterdam: John Benjamins.

Cameron, L. (2003) *Metaphor in Educational Discourse*. London: Continuum.

Cameron, L. and Deignan, A. (2003) Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, 18 (3): 149–60.

Campbell, D., McDonnell, C., Meinardi, M. and Richardson, B. (2007) The need for a speech corpus. *ReCALL*, 19 (1): 3–20.

Campoy-Cubillo, M., Bellés-Fortuño, B. and Gea-Valor, M. (eds) (2010) *Corpus-Based Approaches to English Language Teaching*. London: Continuum.

Candlin, C. (ed.) (2002) *Research and Practice in Professional Discourse*. City University of Hong Kong: City University of Hong Kong Press.

Candlin, C. (2007) Respondent in public lecture 'Press Talk: the UK press and refugees'. City University of Hong Kong, 26 February 2007.

Candlin, C., Bruton, C., Leather, J. and Woods, G. (1981) Designing modular materials for communicative language learning; an example: doctor–patient communication skills. In L. Selinker et al. (eds), *English for Academic and Technical Purposes*, pp. 105–33. Rowley, Mass.: Newbury House.

Candlin, C. and Sarangi, S. (2004) Making applied linguistics matter. *Journal of Applied Linguistics*, 1 (1): 1–8.

Candlin, S. (2006) Constructing knowledge, understanding and meaning between patients and nurses. In M. Gotti and F. Salager-Meyer (eds) *Advances in Medical Discourse Analysis*: *Oral and Written Contexts*, pp. 65–86. Bern: Peter Lang.

Carretero González, M. and Hidalgo Tenorio, E. (2005) For Every Man Hath Business and Desire. Or what modality can do for Hamlet. In Perez Basanta et al. (eds), *Towards an Understanding of the English Language: Past, Present and Future*, pp. 1–7. University of Granada.

Carter, R. (1987) *Vocabulary*. London: Allen and Unwin.

Carter, R. (2004) *Language and Creativity: the Art of Common Talk*. London: Routledge.

Carter, R. and Adolphs, S. (2008) Linking the verbal and the visual: new directions for corpus linguistics. *Language and Computers*, 64: 275–91.

Carter, R. and McCarthy, M. (eds) (1988) *Vocabulary and Language Teaching*. London: Longman.

Carter, R. and McCarthy, M. (1995) Grammar and the spoken language. *Applied Linguistics*, 16 (2): 141–58.

Carter, R. and McCarthy, M. (1997) *Exploring Spoken English*. Cambridge: Cambridge University Press.

Carter, R. and McCarthy, M. (2004) Talking, creating: interactional language, creativity and context. *Applied Linguistics*, 25 (1): 62–88.

Carter, R. and McCarthy, M. (2006) *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N. and Volanschi, A. (2011) Designing a learner translator corpus for training purposes. In N. Kübler (ed.), *Language Corpora, Teaching and Resources: from Theory to Practice*, pp. 200–21. Bern: Peter Lang.

Čermák, F. (2003) Source materials for dictionaries. In P. van Sterkenberg (ed.), *A Practical Guide to Lexicography*, pp. 18–25. Amsterdam: John Benjamins.

Chafe, W. (1970) *Meaning and the Structure of Language*. Chicago: The University of Chicago Press.

Chafe, W. (1992) The importance of corpus linguistics to understanding the nature of language. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp.79–97. Berlin: Mouton de Gruyter.

Chambers, A. (2010) What is data-driven learning? In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 345–58. London: Routledge.

Chambers, J.K. (2000) World enough and time: global enclaves of the near future. *American Speech*, 75: 285–7.

Chang, C.-F. and Kuo, C.-H. (2011) A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, 30: 222–34.

Channell, J. (2000) Corpus-based analysis of evaluative lexis. In S. Hunston and G. Thompson (eds), *Evaluation in Text: Authorial Stance and the Construction of Discourse*, pp. 38–55. Oxford: Oxford University Press.

Charles, M. (2003) 'This mystery…': a corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes*, 2: 313–26.

Charles, M. (2006) The construction of stance in reporting clauses: a cross-disciplinary study of theses. *Applied Linguistics*, 27: 492–518.

Charles, M. (2007) Reconciling top-down and bottom-up approaches to graduate writing: using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6 (4): 289–302.

Charles, M., Pecorari, D. and Hunston, S. (eds) (2009) *Academic Writing: at the Interface of Corpus and Discourse*. London: Continuum.

Charteris-Black, J. (2004) *Corpus Approaches to Critical Metaphor Analysis*. New York: Palgrave Macmillan.

TOOK

Cheng, W. (2004) //→ did you // from the mini<u>BAR</u>//. What is the practical relevance of a corpus-driven language study to practitioners in Hong Kong's hotel industry? In U. Connor and T. Upton (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 141–66. Amsterdam: John Benjamins.

Cheng, W. (2009) *Income/interest/net*. Using internal criteria to determine the aboutness of a text. In K. Aijmer (ed.), *Corpora and Language Teaching*, pp. 157–77. Amsterdam: John Benjamins.

Cheng, W., Greaves, C., Sinclair, J. McH. and Warren, M. (2008a) Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics* 30 (2): 236–52.

Cheng, W., Greaves, C. and Warren, M. (2008b) *A Corpus-Driven Study of Discourse Intonation. The Hong Kong Corpus of Spoken English (Prosodic)*. Amsterdam: John Benjamins.

Cheng, W. and Warren, M. (2008) //→ ONE country two SYStems// The discourse intonation patterns of word associations. In A. Ädel and R. Reppen (eds), *Corpora and Discourse: the Challenges of Different Settings*, pp. 135–53. Amsterdam: John Benjamins.
Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton.
Chuang, F.-Y, and Nesi, H. (2006) An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora* 1 (2): 251–71.
Chujo, K., Utiyama, M. and Nishigaki, C. (2007) Towards building a usable corpus collection for the ELT classroom. In E. Hidalgo, L. Quereda and J. Santana (eds), *Corpora in the Foreign Language Classroom*, pp. 47–69. Amsterdam: Rodopi.
Cobb, T. (1997) Is there any measurable learning from hands-on concordancing? *System*, 25: 301–15.
Cobb, T. (1999) Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12 (4): 345–60.
COBUILD (1987) *Collins Cobuild English Dictionary for Advanced Learners*. Glasgow: HarperCollins Publishers.
COBUILD (1995) *Collins Cobuild English Dictionary for Advanced Learners*, 2nd edn. Glasgow: HarperCollins Publishers.
Coccetta, F. (2011) Multimodal functional-notional concordancing. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 121–38. London: Continuum.
Coffin, C., Hewings, A. and O'Halloran, K. (eds) (2004) *Applying English Grammar: Functional and Corpus Approaches*. The Open University: Arnold.
Coffin, C. and O'Halloran, K. (2005) FINDING THE GLOBAL GROOVE: Theorising and analyzing dynamic reader positioning using APPRAISAL, corpus, and a concordancer. *Critical Discourse Studies*, 2 (2): 143–63.
Collentine, J. and Asención-Delaney, Y. (2010) A corpus-based analysis of the discourse functions of *Ser/Estar* + adjective in three levels of Spanish as FL learners. *Language Learning*, 60 (2): 409–45.
Collins English Thesaurus (1998) London: HarperCollins.
Coniam, D. (1997a) A practical introduction to corpora in a teacher training language awareness programme. *Language Awareness*, 6 (4): 199–207.
Coniam, D. (1997b) A preliminary enquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, 14 (2–4): 15–33.
Connor, U. (2004) Introduction. Special issue on Contrastive Rhetoric in EAP. *Journal of English for Academic Purposes*, 3 (4): 271–6.
Connor, U. (2011) *Intercultural Rhetoric in the Writing Classroom*. Ann Arbor, Mich.: University of Michigan Press.
Connor, U. and Upton, T. (eds) (2002) *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*. Amsterdam: Rodopi.
Connor, U. and Upton, T. (eds) (2004a) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins.
Connor, U. and Upton, T. (eds) (2004b) *Applied Corpus Linguistics: a Multidimensional Perspective*. Amsterdam: Rodopi.
Conrad, S. (1999) The importance of corpus-based research for language teachers. *System*, 27: 1–18.
Cook, G. (1998) The uses of reality: a reply to Ronald Carter. *ELT Journal*, 52 (1): 57–63.
Cook, G. (2001) 'The philosopher pulled the lower jaw of the hen.' Ludicrous invented sentences in language teaching. *Applied Linguistics*, 22 (3): 366–87.
Cook, G. and Seidlhofer, B. (eds) (1995) *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press.

Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes*, 23: 397–423.
Cortes, V. (2007) A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3 (1): 43–57.
Costa, P.T. Jr, Terracciano, A. and McCrae, R. (2001) Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, 81 (2): 322–31.
Cotterill, J. (2001) Domestic discord, rocky relationships: semantic prosodies in representations of marital violence in the courtroom. *Discourse and Society*, 12 (3): 315–36.
Cotterill, J. (2003) *Language and Power in Court. A Linguistic Analysis of the OJ Simpson Trial*. Basingstoke: Palgrave.
Cotterill, J. (2004) Collocation, connotation, and courtroom semantics: lawyer's control of witness testimony through lexical negotiation. *Applied Linguistics*, 25 (4): 513–37.
Cotterill, J. (2010) How to use corpus linguistics in forensic linguistics. In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 578–90. London: Routledge.
Coulthard, M. (1993) Beginning the study of forensic texts: corpus, concordance, collocation. In M. Hoey (ed.), *Data, Description, Discourse*, pp. 86–97. London: HarperCollins.
Coulthard, M. (1994) On the use of corpora in the analysis of forensic texts. *Forensic Linguistics*: *The International Journal of Speech, Language and the Law*, 1 (1): 27–43.
Coulthard, M. (1995) Explorations in applied linguistics 3: forensic stylistics. In G. Cook and B. Seidlhofer (eds), *Principle and Practice in Applied Linguistics*, pp. 229–43. Oxford: Oxford University Press.
Coulthard, M. (1996) The official version. Audience manipulation in police records of interviews with suspects. In C. Caldas-Coulthard and M. Coulthard (eds), *Texts and Practices*, pp. 166–78. London: Routledge.
Coulthard, M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25 (4): 431–47.
Coulthard, M. and Johnson, A. (2007) *An Introduction to Forensic Linguistics: Language in Evidence*. London: Continuum.
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
Coupland, N. (1998) What is sociolinguistic theory? *Journal of Sociolinguistics*, 2 (1): 110–17.
Coupland, N. (2001) Sociolinguistic theory and social theory. In N. Coupland, S. Sarangi and C. Candlin (eds), *Sociolinguistics and Social Theory*, pp. 1–26. Essex: Pearson.
Cowie, A.P. (ed.) (1998) *Phraseology: theory, analysis and applications*. Oxford Studies in Lexicography and Lexicology. Oxford: Clarendon Press.
Cowie, A. and Howarth, P. (1996) Phraseological competence and written proficiency. In G. Blue and R. Mitchell (eds), *Language and Education: British Studies in Applied Linguistics, 11*, pp. 80–93. Clevedon: Multilingual Matters.
Coxhead. A. (2000) A new academic wordlist. *TESOL Quarterly*, 34 (2): 213–38.
Coxhead, A. and Byrd, P. (2007) Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16 (3): 129–47.
Cresswell, A. (2007) Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo et al. (eds), *Corpora in the Foreign Language Classroom*, pp. 267–87. Amsterdam: Rodopi.
Croft, W. (2001) *Radical Construction Grammar*. Oxford: Oxford University Press.
Crowdy, S. (1993) Spoken corpus design. *Literary and Linguistic Computing*, 8: 259–65.

Culpepper, J. (2002) Computers, language and characterization: an analysis of six characters in Romeo and Juliet. In U. Melander-Marttala, C. Östman and M. Kytö (eds), *Conversation in Life and Literature*, pp. 11–33. Uppsala: Universitetstryckeriet.

Curado Fuentes, A. (2002) Exploitation and assessment of a business English corpus through language learning tasks. *ICAME Journal*, 26: 5–32.

Cutting, J. (ed.) (2007) *Vague Language Explored*. Basingstoke: Palgrave Macmillan.

Cutting, J. (2008) *Pragmatics and Discourse*, 2nd edn. London: Routledge.

Dagneaux, E., Denness, S. and Granger, S. (1998) Computer-aided error analysis. *System*, 26: 163–74.

Davies, M. (n.d.) *BYU-BNC: British National Corpus*, http://corpus.byu.edu/bnc/x.asp

De Beaugrande, R. (1997) *New Foundations for a Science of Text and Discourse*. Norwood, NJ: Ablex Publishing Corporation.

De Beaugrande, R. (2002) Linguistics, sociolinguistics, and corpus linguistics: ideal language versus real language. *Journal of Sociolinguistics*, 3 (1): 128–38.

De Cock, S. (2000) Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair and M. Hundt (eds), *Corpus Linguistics and Linguistic Theory*, pp. 51–88. Amsterdam: Rodopi.

De Cock, S. (2011) Preferred patterns of use of positive and negative evaluative adjectives in native and learner speech: an ELT perspective. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 198–212. London: Continuum.

De Mönnink, I. (1997) Using corpus and experimental data: a multimethod approach. In M. Ljung (ed.) *Corpus-Based Studies in English*, pp. 227–44. Amsterdam: Rodopi.

Deane, P. and Quinlan, T. (2010) What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2 (2): 151–77.

Deignan, A. (2005) *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Douglas, F. (2003) The Scottish corpus of texts and speech: problems of corpus design. *Literary and Linguistic Computing*, 18 (1): 23–37.

Duguid, A. (2007) Soundbiters bit. Contracted dialogistic space and the textual relations of the No. 10 team analysed through corpus-assisted discourse studies. In N. Fairclough, G. Cortese and P. Ardizzone (eds), *Discourse and Contemporary Social Change*, pp. 73–94. Frankfurt am Main: Peter Lang.

Duguid, A. (2009) Loud signatures. Comparing evaluative discourse styles. In U. Römer and R. Schulze (eds), *Exploring the Lexis–Grammar Interface*, pp. 289–315. Amsterdam: John Benjamins.

Dulay, H. and Burt, M. (1975) Creative construction in second language learning and teaching. In M. Burt and H. Dulay (eds), *On TESOL '75: New Directions in Second Language Learning, Teaching and Bilingual Education*. Washington, DC: TESOL.

Durrant, P. (2009) Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28: 157–69.

Durrant, P. and Mathews-Aydınlı, J. (2011) A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30 (1): 58–72.

Ellis, N. (1994) *Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, N., 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., Simpson-Vlach, R. and Maynard, C. (2008) Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly*, 42 (3): 375–96.

Engeström, Y. and Middleton, D. (eds) (1996) *Cognition and Communication at Work*. Cambridge: Cambridge University Press.

Erman, B. (2007) Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*, 12 (1): 25–53.
Erman, B. and Warren, B. (2000) The idiom principle and the open choice principle. *Text*, 20 (1): 29–62.
Fairclough, N. (1995) *Critical Discourse Analysis*. London: Longman.
Fairclough, N. (2000) *New Labour, New Language?* London: Routledge.
Fairclough, N. (2003) *Analysing Discourse*. London: Routledge.
Fan, M. and Xu, X. (2002) An evaluation of an online bilingual corpus for the self-learning of legal English. *System*, 30 (1): 47–63.
Farr, F. (2003) Engaged listenership in spoken academic discourse: the case of student–tutor meetings. *Journal of English for Academic Purposes*, 2 (2): 67–85.
Farr, F. (2008) Evaluating the use of corpus-based instruction in a language teacher education context: perspectives from the users. *Language Awareness*, 17 (1): 25–43.
Farr, F. (2010) *The Discourse of Teaching Practice Feedback. A Corpus-Based Investigation of Spoken and Written Modes*. London: Routledge.
Farr, F. and O'Keeffe, A. (2002) *Would* as a hedging device in an Irish context: an intra-varietal comparison of institutionalized spoken interaction. In R. Reppen, S. Fitzpatrick and D. Biber (eds), *Using Corpora to Explore Linguistic Variation*, pp. 25–48. Amsterdam: John Benjamins.
Farr, F. and O'Keeffe, A. (eds) (2011) *International Journal of Corpus Linguistics*. Special issue on Teacher Education, 16 (3).
Feak, C., Reinhart, S., and Rohlck, T. (2009) *Academic Interactions. Communicating on Campus*. Ann Arbor, Mich.: University of Michigan Press.
Ferguson, G. (2001) If you pop over there: a corpus-based study of conditionals in medical discourse. *English for Specific Purposes*, 20 (2): 61–82.
Fillmore, C. (1968) The case for case. In E. Bach and T.T. Harms (eds), *Universals in Linguistic Theory*, pp. 1–88. New York: Holt, Rinehart and Winston.
Fillmore, C. (1992) 'Corpus linguistics' or 'computer-aided armchair linguistics'. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 35–60. Berlin: Mouton de Gruyter.
Fillmore, C. (2001) Armchair linguistics vs. corpus linguistics revisited. Keynote paper presented at ICAME 2001: Future Challenges in Corpus Linguistics. Louvain-la-Neuve, Belgium, 16–20 May.
Fillmore, C. (2006) Frame semantics. In D. Geeraerts (ed.), *Cognitive Linguistics*, pp. 373–400. Berlin: Mouton de Gruyter.
Firth, J. R. (1957) *Papers in Linguistics*. London: Oxford University Press.
Fischer-Starcke, B. (2009) Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*. A corpus-stylistic analysis. *International Journal of Corpus Linguistics*, 14 (4): 492–523.
Fischer-Starcke, B. (2010) *Corpus Linguistics in Literary Analysis. Jane Austen and her Contemporaries*. London: Continuum.
Fish, S. (1996) What is stylistics and why are they saying such terrible things about it? In J.J. Weber (ed.), *The Stylistics Reader*, pp. 94–116. London: Arnold.
Fletcher, W. (2002) Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (eds), *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*, pp. 191–205. Amsterdam: Rodopi.
Flowerdew, J. (2002a) Computer-assisted analysis of language learner diaries. A qualitative application of word frequency and concordancing software. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 231–43. Amsterdam: Rodopi.

Flowerdew, J. (ed.) (2002b) *Academic Discourse*. London: Longman.
Flowerdew, J. (2003a) Register-specificity of signalling nouns in discourse. In P. Leistyna and C. Meyer (eds), *Corpus Analysis: Language Structure and Language Use*, pp. 35–46. Amsterdam: Rodopi.
Flowerdew, J. (2003b) Signalling nouns in discourse. *English for Specific Purposes*, 22: 329–46.
Flowerdew, J. (2009) Corpora in language teaching. In M. Long and C. Doughty (eds), *The Handbook of Language Teaching*, pp. 327–50. London: Wiley-Blackwell.
Flowerdew, J. and Forest, R. (2009) Schematic structure and lexico-grammatical realization in corpus-based genre analysis: The case of 'Research' in the PhD literature review. In M. Charles et al. (eds), *Academic Writing: at the Interface of Corpus and Discourse*, pp. 15–36. London: Continuum.
Flowerdew, L. (1998) Corpus linguistic techniques applied to textlinguistics. *System*, 26: 541–52.
Flowerdew, L. (2000) Investigating referential and pragmatic errors in a learner corpus. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 145–54. Frankfurt am Main: Peter Lang.
Flowerdew, L. (2003) A combined corpus and systemic-functional analysis of the Problem–Solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly*, 37 (3): 489–511.
Flowerdew, L. (2004) The argument for using English specialised corpora to understand academic and professional language. In U. Connor and T. Upton (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 11–33. Amsterdam: John Benjamins.
Flowerdew, L. (2005) An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24: 321–32.
Flowerdew, L. (2008a) *Corpus-Based Analyses of the Problem–Solution Pattern*. Amsterdam: John Benjamins.
Flowerdew, L. (2008b) Corpus linguistics for academic literacies mediated through discussion activities. In D. Belcher and A. Hirvela (eds), *The Oral-Literate Connection. Perspectives on L2 Speaking, Writing, and Other Media Interactions*, pp. 268–87. Ann Arbor, Mich.: University of Michigan Press.
Flowerdew, L. (2009) Applying corpus linguistics to pedagogy: a critical evaluation. *International Journal of Corpus Linguistics*, 4 (3): 393–417.
Flowerdew, L. (2010) Using corpora for writing instruction. In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 444–57. London: Routledge.
Flowerdew, L. (2011a) ESP and corpus studies. In D. Belcher, A. Johns and B. Paltridge (eds), *New Directions in English for Specific Purposes Research*, pp. 222–51. Ann Arbor, Mich.: University of Michigan Press.
Flowerdew, L. (2011b) Corpus-based discourse analysis. In J.P. Gee and M. Handford (eds), *Routledge Handbook of Discourse Analysis*, pp. 174–87. London: Routledge.
Flowerdew, L. (in press, 2012a) Corpora and the classroom: an applied linguistic perspective. In K. Hyland, M. H. Chau, and M. Handford (eds), *Corpora in Applied Linguistics: Current Approaches and Future Directions*. London: Continuum.
Flowerdew, L. (in press, 2012b) Exploiting a corpus of business letters from a phraseological, functional perspective. *ReCALL*.
Foucou, P.-Y. and Kübler, N. (2000) A web-based environment for teaching technical English. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 65–71. Frankfurt am Main: Peter Lang.

Foucou, P.-Y. and Kübler, N. (2003) Teaching English verbs with bilingual corpora: examples in the field of computer science. In S. Granger et al. (eds) *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*, pp. 185–206. Amsterdam: Rodopi.

Francis, G. (1991) Nominal group heads and clause structure. *Word*, 42 (2): 145–56.

Francis, G. (1993) A corpus-driven approach to grammar – principles, methods and examples. In M. Baker, G. Francis and E. Tognini Bonelli (eds), *Text and Technology*, pp. 137–56. Amsterdam: John Benjamins.

Francis, N. (1992) Language corpora B.C. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 17–32. Berlin: Mouton de Gruyter.

Frankenberg-Garcia, A. (2004) Lost in parallel concordances. In G. Aston et al. (eds), *Corpora and Language Learners*, pp. 213–29. Amsterdam: John Benjamins.

Frankenberg-Garcia, A. (2005a) Pedagogical uses of monolingual and parallel concordances. *ELT Journal*, 59 (3): 189–98.

Frankenberg-Garcia, A. (2005b) A peek into what today's language learners as researchers actually do. *International Journal of Lexicography*, 18 (3): 335–55.

Frankenberg-Garcia, A. (2009) Are translations longer than source texts? A corpus-based study of explicitation. In A. Beeby et al. (eds), *Corpus Use and Translating*, pp. 47–58. Amsterdam: John Benjamins.

Frankenberg-Garcia, A. (2010) Raising teachers' awareness of corpora. *Language Teaching*, doi: 10.1017/S0261444810000480, published online by Cambridge University Press.

Frankenberg-Garcia, A., Flowerdew, L. and Aston, G. (eds) (2011) *New Trends in Corpora and Language Learning*. London: Continuum.

Freedman, A. and Medway, P. (eds) (1994) *Genre and the New Rhetoric*. London: Taylor and Francis.

Fries, C. (1952) *The Structure of English: an Introduction to the Construction of Sentences*. New York: Harcourt Brace.

Friginal, E. (2009) *The Language of Outsourced Call Centers. A Corpus-Based Study of Cross-Cultural Interaction*. Amsterdam: John Benjamins.

Fung, L. and Carter, R. (2007) Discourse markers and spoken English: native and learner use in pedagogic settings. *Applied Linguistics*, 23 (3): 410–39.

Gabrielatos, C. (2005) Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ*, 8 (4): 1–37.

Garside, R., Leech, G. and McEnery, A. (eds) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpus*. Harlow: Addison Wesley Longman.

Gaskell, D. and Cobb, T. (2004) Can learners use concordance feedback for writing errors? *System*, 32: 301–19.

Gavioli, L. (2002) Some thoughts on the problem of representing ESP through small corpora. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 293–303. Amsterdam: Rodopi.

Gavioli, L. (2005) *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Gee, J.P. (2001) *An Introduction to Discourse Analysis*. London: Routledge.

Gee, J.P. and Handford, M. (eds) (2011) *Routledge Handbook of Discourse Analysis*. London: Routledge.

Ghadessy, M. (2003) Comments on Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd and Marie Helt's 'Speaking and Writing in the University: a Multidimensional Comparison'. *TESOL Quarterly*, 37 (1): 147–50.

Ghadessy, M. Henry, A., and Roseberry, R. (eds) (2001) *Small Corpus Studies and ELT*. Amsterdam: John Benjamins.

Gillard, P. and Gadsby, A. (1998) Using a learners' corpus in compiling ELT dictionaries. In S. Granger (ed.), *Learner English on Computer*, pp. 159–71. London: Longman.
Gilquin, G. (2000/2001) The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast*, 3 (1): 95–123.
Gilquin, G. (2003) Causative *get* and *have*. *Journal of English Linguistics*, 31 (2): 125–48.
Gilquin G. (2006) The place of prototypicality in corpus linguistics: causation in the hot seat. In S. Gries and A. Stefanowitsch (eds), *Corpora in Cognitive Linguistics*, pp. 159–91. Berlin: Mouton de Gruyter.
Gilquin, G. (2008) Hesitation markers among EFL learners: pragmatic deficiency or difference? In J. Romero-Trillo (ed.) *Pragmatics and Corpus Linguistics*, pp. 119–49. New York: Mouton de Gruyter.
Gilquin, G. (2010) *Corpus, Cognition and Causative Constructions*. Amsterdam: John Benjamins
Gilquin, G., Granger, S. and Paquot, M. (2007) Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6 (4): 319–35.
Gilquin, G. and Gries, S. 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. (eds) (2008) *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi.
Gilquin, G. and Paquot, M. (2008) Too chatty: learner academic writing and register variation. *English Text Construction*, 1 (1): 41–61.
Giroux, H. (1992) *Border Crossings: Cultural Workers and the Politics of Education*. London: Routledge.
Gisborne, N. (2000) Relative clauses in Hong Kong English. In Bolton, K. (ed.), *Hong Kong English: Autonomy and Creativity*. Special issue of *World Englishes*, 19: (3): 357–72.
Glaister, L. (1992) *Digging to Australia*. London: Secker and Warburg.
Glaister, L. (1995) *Buddeln bis Australien*, translated by H. C. Oeser. Zurich: Diogenes.
Gledhill, C. (2000) The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19: 115–35.
Gnutzmann, C. (2009) Language for specific purposes vs. general language. In K. Knapp and B. Seidlhofer in cooperation with H. G. Widdowson (eds), *Handbook of Foreign Language Communication and Learning*, pp. 517–44. Amsterdam: Mouton de Gruyter.
Goffman, E. (1974) *Frame Analysis*. New York: Harper and Row.
Goffman, E. (1981) *Forms of Talk*. Philadelphia, Pa: University of Pennsylvania Press.
Goldberg, L. R. (1981) Language and individual differences: the search for universals in personality lexicons. In L. Wheeler (ed.), *Review of Personality and Social Psychology*, 2: 141–65.
Gouverneur, C. (2008) The phraseological patterns of high-frequency verbs in advanced English for general purposes. In F. Meunier and S. Granger (eds), *Phraseology in Foreign Language Teaching and Learning*, pp. 223–43. Amsterdam: John Benjamins.
Graddol, D. (2006) *English Next*. British Council.
Granger, S. (1993) The international corpus of learner English. In J. Aarts, P. de Haan and N. Oostdijk (eds), *English Language Corpora*, pp. 57–69. Amsterdam: Rodopi.
Granger, S. (1996) From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In K. Aijmer et al. (eds), *Languages in Contrast*, pp. 37–51. Lund: Lund University Press.
Granger, S. (ed.) (1998a) *Learner English on Computer*. London: Longman.
Granger, S. (1998b) Prefabricated patterns in advanced EFL writing: collocations and formulae, In A. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*, pp. 145–60. Oxford Studies in Lexicography and Lexicology. Oxford: Clarendon Press.

Granger, S. (1998c) The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (ed.), *Learner English on Computer*, pp. 3–18. London: Longman.

Granger, S. (2002) A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 3–33. Amsterdam: John Benjamins.

Granger, S. (2004) Computer learner corpus research: current status and future prospects. In U. Connor and T. Upton (eds), *Applied Corpus Linguistics: a Multidimensional Perspective*, pp. 123–45. Amsterdam: Rodopi.

Granger, S. (2009) The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (ed.), *Corpora and Language Teaching*, pp. 13–32. Amsterdam: John Benjamins.

Granger, S., Hung, J. and Petch-Tyson, S. (eds) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Granger, S. and Meunier, F. (2008) Phraseology in language learning and teaching: where to from here? In S. Granger and F. Meunier (eds), *Phraseology in Foreign Language Learning and Teaching*, pp. 247–51. Amsterdam: John Benjamins.

Greaves, C. (2009) *ConcGram 1.0: a Phraseological Search Engine*. Amsterdam: John Benjamins.

Green, E. and Peters, P. (1991) The Australian corpus project and Australian English. *ICAME Journal*, 15: 37–53.

Greenbaum, S. (ed.) (1996) *Comparing English Worldwide: the International Corpus of English*. Oxford: Clarendon Press.

Gries, S. Th. (2006) Corpus-based methods and cognitive semantics: the many senses of *to run*. In S. Th. Gries and A. Stefanowitsch (eds), *Corpora in Cognitive Linguistics*, pp. 57–99. Berlin: Mouton de Gruyter.

Gries, S. Th. and Stefanowitsch, A. (eds) (2006) *Corpora in Cognitive Linguistics*. Berlin: Mouton de Gruyter.

Grundmann, R. and Krishnamurthy, R. (2010) The discourse of climate change: a corpus-based approach. *Critical Approaches to Discourse Analysis across Disciplines*, 4 (2): 125–46.

Gu, Yueguo (2002) Towards an understanding of workplace discourse. A pilot study for compiling a spoken Chinese corpus of situated discourse. In C. Candlin (ed.), *Research and Practice in Professional Discourse*, pp. 137–186. Hong Kong: City University of Hong Kong Press.

Gu, Yueguo (2006) Multimodal text analysis: a corpus linguistic approach to situated discourse. *Text and Talk*, 26 (2): 127–67.

Gu, Yueguo (n.d.) Compiling a spoken Chinese corpus of situated discourse (http://ling.cass.cn/dangdai.corpus.htm). Retrieved 12 December 08.

Gumperz, J. (1992) Contextualisation and understanding. In A. Duranti and C. Goodwin (eds), *Rethinking Context: Language as an Interactive Phenomenon*, pp. 229–52. Cambridge: Cambridge University Press.

Haarman, L. and Lombardo, L. (eds) (2009) *Evaluation and Stance in War News*. London: Continuum.

Hafner, C. (2008) Designing, implementing and evaluating an online resource for professional legal communication skills. Unpublished doctoral thesis, Macquarie University, Sydney. Available at http://personal.cityu.edu.hk/~elhafner/reserch/hafnerphdthesis.html.

Hafner, C. and Candlin, C. (2007) Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, 6 (4): 303–18.

Hahn, A. (2000) Grammar at its best: the development of a rule- and corpus-based grammar of English tenses. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 193–205. Frankfurt am Main: Peter Lang.

Hall, D. and Beggs, E. (1998) Defining learner autonomy. In W. Renandya and G. Jacobs (eds), *Learners and Language Learning*, pp. 26–39. SEAMEO-RELC Anthology Series 39.

Halliday, M.A.K. (1991) Corpus studies and probabilistic grammars. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 30–43. Amsterdam: Rodopi.

Halliday, M.A.K. (1992) Language as system and language as instance: the corpus as a theoretical construct. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 61–77. Berlin: Mouton de Gruyter.

Halliday, M.A.K. (1993) Quantitative studies and probabilities in grammar. In M. Hoey (ed.), *Data, Description, Discourse*, pp. 1–25. London: HarperCollins.

Halliday, M.A.K. (2004) The spoken language corpus: a foundation for grammatical theory. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 11–38. Amsterdam: Rodopi.

Halliday, M.A.K. (2008) *Complementarities in Language*. Beijing: Commercial Press.

Handford, M. (2010a) What can a corpus tell us about specialist genres? In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Linguistics*, pp. 255–69. London: Routledge.

Handford, M. (2010b) *The Language of Business Meetings*. Cambridge: Cambridge University Press.

Handford, M. and Koester, A. (2010) 'It's not rocket science': metaphors and idioms in conflictual business meetings. *Text and Talk*, 30 (1): 27–51.

Handford, M. and Matous, P. (2011) Lexicogrammar in the international construction industry: a corpus-based study of Japanese-Hong-Kongese on-site interactions. *English for Specific Purposes*, 30 (2): 87–100.

Hanks, P. (2002) Mapping meaning onto use. In M.-H. Corréard (ed.), *Lexicography and Natural Language Processing: a Festschrift in Honour of B.T.S. Atkins*. Euralex.

Hanks, P. (2009) The impact of corpora on dictionaries. In P. Baker (ed.), *Contemporary Corpus Linguistics*, pp. 214–36. London: Continuum.

Hardt-Mautner, G. (1995) Only connect: critical discourse analysis and corpus linguistics. UCREL Technical Paper 6. Lancaster: University of Lancaster.

Hargreaves, P. (2000) Collocation and testing. In M. Lewis (ed.) *Teaching Collocation*, pp. 205–23. Hove: Language Teaching Publications.

Harris, Z. (1954) Distributional structure. *Word*, 10 (2–3): 146–62.

Hartmann, R.R.K. (2001) *Teaching and Researching Lexicography*. London: Pearson.

Harvey, K. and Adolphs, S. (2011) Discourse and healthcare. In J.P. Gee and M. Handford (eds), *Routledge Handbook of Discourse Analysis*, pp. 470–81. London: Routledge.

Harvey, K., Churchill, D., Crawford, P., Brown, B., Mullany, L. Macfarlane, A. and McPherson, A. (2008) Health communication and adolescents: what do their e-mails tell us? *Family Practice*, 25 (4): 304–11.

Harwood, N. (2005a) 'We do not seem to have a theory…The theory I present here attempts to fill this gap': inclusive and exclusive pronouns in academic writing. *Applied Linguistics*, 26 (3): 343–75.

Harwood, N. (2005b) A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics*, 37: 1207–31.

Hasselgren, H. (2002) Learner corpora and language testing: smallwords as markers of learner fluency. In S. Granger, J. Hung and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 143–73. Amsterdam: John Benjamins.

Hatim, B. (2001) *Teaching and Researching Translation*. New York: Longman.
Hatzitheodorou, A.-M. and Mattheoudakis, M. (2011) The impact of culture on the use of exponents as persuasive devices: the case of GRICLE and English native speaker corpora. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 229–46. London: Continuum.
Hawkey, R. and Barker, F. (2004) Developing a common scale for the assessment of writing. *Assessing Writing*, 9 (2): 122–59.
He, A. and Kennedy, G. (1999) Successful turn-bidding in English conversation. *International Journal of Corpus Linguistics*, 4 (1): 1–27.
HeadTalk project (http://www.ncess.ac.uk/research/sgp/headtalk/). Retrieved 12 Dec. 08.
Heid, U. (2009) Corpus linguistics and lexicography. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook*, vol. 1, pp. 131–53. New York: Mouton de Gruyter.
Heffer, C. (2005) *The Language of Jury Trial: a Corpus-Aided Analysis of Legal–Lay Discourse*. London: Palgrave Macmillan.
Henderson, A. and Barr, R. (2010) Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research*, 2 (2): 245–64.
Hewings, A., Coffin, C. and North, S. (2009) E-conferencing: corpus and discourse insights. In M. Charles et al. (eds), *Academic Writing: at the Interface of Corpus and Discourse*, pp. 129–51. London: Continuum.
Hewings, M. and Hewings, A. (2002) 'It is interesting to note that…': a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, 21 (4): 367–83.
Heyvaert, L. and Laffut, A. (2008) Corpora in the teaching of English in Flemish secondary schools: current situation and future perspectives. Paper presented at the 8th Teaching and Language Corpora Conference, Lisbon, Portugal, 3–6 July.
Hidalgo, E., Quereda, L. and Santana, J. (eds) (2007) *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.
Highfield, R. (4–5 Dec. 2004) In her own words. *The Daily Telegraph*, p. A31.
Hilton Hubbard, E. (1997) Errors in court: a forensic application of error analysis. In H. Kniffka et al. (eds), *Recent Developments in Forensic Linguistics*, pp. 123–39. Frankfurt am Main: Peter Lang.
Hoey, M. (ed.) (1993) *Data, Description, Discourse*. London: HarperCollins.
Hoey, M. (1996) Cohesive words: a paper of consequence. In J. Svartvik (ed.) *Words: Proceedings of an International Symposium*, Lund, 25–26 August 1995, Stockholm.
Hoey, M. (1997) From concordance to text structure: new uses for computer corpora. In J. Melia and B. Lewandowska (eds), *Practical Applications in Language Corpora*, pp. 2–23. University of Lodz, Poland.
Hoey, M. (1998) Some text properties of certain nouns. In T. McEnery and S. Botley (eds), *Proceedings of the Colloquium on Discourse Anaphora and Reference Resolution*. Lancaster: University of Lancaster.
Hoey, M. (2001). *Textual Interaction: an Introduction to Written Discourse Analysis*. London: Routledge.
Hoey, M. (2004a) Textual colligation: a special kind of lexical priming. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 171–94. Amsterdam: Rodopi.
Hoey, M. (2004b) The textual priming of lexis. In G. Aston et al. (eds), *Corpora and Language Learners*, pp. 21–41. Amsterdam: John Benjamins.
Hoey, M. (2005) *Lexical Priming: a New Theory of Words and Language*. London: Routledge.

Hoey, M. (2007) Lexical priming and literary creativity. In M. Hoey et al. (eds), *Text, Discourse and Corpora*, pp. 7–29. London: Continuum.

Hoey, M. and Brook O'Donnell, M. (2008) Lexicography, grammar and texual position. *International Journal of Lexicography*, 21 (3): 293–309.

Hoey, M., Mahlberg, M., Stubbs, M. and Teubert, W. (eds) (2007) *Text, Discourse and Corpora*. London: Continuum.

Holmes, J. (1978) The future of translation theory: a handful of theses. Paper presented at the International Symposium on Achievements in the Theory of Translation, 23–30 October, Moscow. Reprinted in Holmes (1994), Translated, pp. 81–91. Amsterdam: Rodopi.

Holmes, J. (2004) Talk at work and 'fitting in': a socio-pragmatic perspective on workplace culture. In G. Wigglesworth (ed.), *Proceedings of Conference on Language Education in Australian and New Zealand Universities 2003*, pp. 95–115. Melbourne: University of Melbourne.

Holmes, J. (2005) When small talk is a big deal: sociolinguistic challenges in the workplace. In M. Long (ed.), *Second Language Needs Analysis*, pp. 344–71. Cambridge: Cambridge University Press.

Holmes, J. and Sigley, R. (2002) What's a word like *girl* doing in a place like this? In P. Peters, P. Collins and A. Smith (eds), *New Frontiers of Corpus Research*, pp. 247–64. Amsterdam: Rodopi.

Hori, M. (2002) Collocational patterns of *-ly* manner adverbs in Dickens. In T. Saito, J. Nakamura and S. Yamazaki (eds), *English Corpus Linguistics in Japan*, pp. 149–63. Amsterdam: Rodopi.

Housen, A. (2002) A corpus-based study of the L2-acquisition of the English verb system. In S. Granger et al. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 77–118. Amsterdam: John Benjamins.

Huddleston, R. and Pullum, G. (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hughes, R. and McCarthy, M. (1998) From sentence to grammar: discourse grammar and English language teaching. *TESOL Quarterly*, 32 (2): 263–87.

Hundt, M. (2009) Global English – global corpora: report on a panel discussion at the 28th ICAME Conference. In A. Renouf and A. Kehoe (eds), *Corpus Linguistics. Refinements and Reassessments*, pp. 451–62. Amsterdam: Rodopi.

Hundt, M. and Biewer, C. (2007) The dynamics of inner and outer circle varieties in the South Pacific and East Asia. In M. Hundt et al. (eds), *Corpus Linguistics and the Web*, pp. 249–69. Amsterdam: Rodopi.

Hundt, M., Nesselhauf, N. and Biewer, C. (eds) (2007) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Hunston, S. (1995) Grammar in teacher education: the role of a corpus. *Language Awarenesss*, 4 (1): 15–29.

Hunston, S. (2001) Colligation, lexis, pattern and text. In M. Scott and G. Thompson (eds), *Patterns of Text*, pp. 13–33. Amsterdam: John Benjamins.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2006) Phraseology and system: a contribution to the debate. In G. Thompson and S. Hunston (eds), *System and Corpus: Exploring Connections*, pp. 55–80. Equinox.

Hunston, S. (2009) Corpus compilation. Collection strategies and design decisions. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics: an International Handbook*, vol. 2, pp. 154–68. Berlin: Mouton de Gruyter.

Hunston, S. (2010) *Corpus Approaches to Evaluation*. London: Routledge.

Hunston, S. and Francis, G. (2000) *Pattern Grammar*. Amsterdam: John Benjamins.

Hunston, S. and Thompson, G. (eds) (2000) *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.

Hüttner, J., Smit, U. and Mehlmauer-Larcher, B. (2009) ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis. *System*, 37 (1): 99–109.

Hyland, K. (1998) *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, K. (2002) *Teaching and Researching Writing*. London: Longman Pearson.

Hyland, K. (2004) Disciplinary interactions: metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing*, 13: 133–51.

Hyland, K. (2005) *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.

Hyland, K. (2008) As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27 (1): 4–21.

Hyland, K. (2009) *Academic Discourse*. London: Continuum.

Hyland, K., Chau, M. H. and Handford, M. (eds) (in press, 2012) *Corpora in Applied Linguistics: Current Approaches and Future Directions*. London: Continuum.

Hyland, K. and Milton, J. (1997) Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6 (2): 183–205.

Hyland, K. and Tse, P. (2004) Metadiscourse in academic writing: a reappraisal. *Applied Linguistics*, 25 (2): 156–77.

Hyland, K. and Tse, P. (2007) Is there an 'academic vocabulary'? *TESOL Quarterly*, 41 (2): 235–53.

Hymes, D. (1971) *On Communicative Competence*. Philadelphia: University of Pennsylvania Press.

Hyon, S. (1996) Genre in three traditions; implications for ESL. *TESOL Quarterly*, 30 (4): 693–722.

Ishii, Y. (2009) Making a list of essential phrasal verbs based on large corpora and phrasal verb dictionaries. In Y. Kawaguchi, M. Minegishi and J. Durand (eds), *Corpus Analysis and Variation in Linguistics*, pp. 121–40. Amsterdam: John Benjamins.

James, G., Davison, R., Cheung, A.H.Y. and Deerwester, S. (1994) *English in Computer Science: a Corpus-Based Lexical Analysis*. Hong Kong: Longman Asia.

Johansson, S. (1991) Times change, and so do corpora. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 305–14. Amsterdam: Rodopi.

Johansson, S. (2009) Some thoughts on corpora and second-language acquisition. In K. Aijmer (ed.), *Corpora and Language Teaching*, pp. 33–44. Amsterdam: John Benjamins.

Johns, T. (1988) Whence and whither classroom concordancing? In T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds), *Computer Applications in Language Learning*, pp. 9–33. Dordrecht: Foris.

Johns, T. (1991) Should you be persuaded: two examples of data-driven learning. *English Language Research Journal*, 4: 1–16.

Johns, T., Lee, H.-C. and Wang, L. (2008) Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Asssisted Language Learning*, 21 (5): 483–506.

Jones, M. and Haywood, S. (2004) Facilitating the acquisition of formulaic sequences: an exploratory study in an EAP context. In N. Schmitt (ed.), *Formulaic Sequences*, pp. 269–91. Amsterdam: John Benjamins.

Jones, M. and Schmitt, N. (2010) Developing materials for discipline-specific vocabulary and phrases in academic seminars. In N. Harwood (ed.), *English Language Teaching Materials. Theory and Practice*, pp. 225–50. Cambridge: Cambridge University Press.

Jucker, A., Schreier, D. and Hundt, M. (2009a) Corpus linguistics, pragmatics and discourse. In A. Jucker et al. (eds), *Corpora: Pragmatics and Discourse*, pp. 3–9. Amsterdam: Rodopi.

Kachru, B. (1986) The power and politics of English. *World Englishes*, 5: 121–40.
Kaltenböck, G. and Mehlmauer-Larcher, B. (2005) Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL*, 17 (1): 65–84.
Kandil, M. and Belcher, D. (2011) ESP and corpus-informed critical discourse analysis: understanding the power of genres of power. In D. Belcher et al. (eds), *New Directions in English for Specific Purposes Research*, pp. 252–70. Ann Arbor, Mich.: University of Michigan Press.
Kanoksilapatham, B. (2005) Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24 (3): 269–92.
Kanoksilapatham, B. (2007) Rhetorical moves in biochemistry research articles. In D. Biber et al. (eds), *Discourse on the Move*, pp. 73–119. Amsterdam: John Benjamins.
Kaszubski, P. (2011) IFA Conc – a pedagogic tool for online concordancing with EFL/EAP learners. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 81–104. London: Continuum.
Kehoe, A. and Gee, M. (2009) A wiki tool for the collaborative study of literary texts. Paper presented in colloquium 'Corpus Linguistics and Literature'. Corpus Linguistics Conference, University of Birmingham, UK, 20 July.
Kelly, T., Nesi, H. and Revell, R. (2000) *EASE Volume One: Listening to Lectures*. University of Warwick.
Kennedy, C. and Miceli, T. (2002) The *CWIC* Project: Developing and using a corpus for intermediate Italian students. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 183–92. Amsterdam: Rodopi.
Kennedy, C. and Miceli, T. (2010) Corpus-assisted creative writing: introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology*, 14 (1): 28–44.
Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman.
Kennedy, G. (2003) Amplifier collocations in the British National Corpus: implications for English language teaching. *TESOL Quarterly*, 37: 477–86.
Kennedy, G. (2005) Collocational patterning with high frequency verbs in the British National Corpus. Paper presented at the American Association of Applied Corpus Linguistics Conference, University of Michigan, Ann Arbor, 12–15 May 2005.
Kenny, D. (1998) Creatures of habit? What translators usually do with words. *Meta*, XLIII (4): 1–9.
Kenny, D. (2000) Translators at play: exploitation of collocational norms in German–English translation. In B. Dodd (ed.), *Working with German Corpora*, pp 143–60. University of Birmingham, UK.
Kettemann, B. and Marko, G. (eds) (2002) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi.
Kettemann, B. and G. Marko (2004) Can the L in TaLC stand for literature? In G. Aston et al. (eds), *Corpora and Language Learners*, pp. 169–93. Amsterdam: John Benjamins.
Kilgariff, A. (2001) Web as corpus. In P. Rayson et al. (eds), *Proceedings of the Corpus Linguistics 2001 Conference*, pp. 342–4. Lancaster: UCREL.
Kilgariff, A. and Grefenstette (2003) Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29 (3): 333–47.
Kilgariff, A. and Rundell, M. (2002) Lexical profiling software and its lexicographic applications: case study. In A. Braasch and C. Povlsen (eds), *Proceedings of the Tenth EURALEX International Congress*. *EURALEX 2002*. Copenhagen: Center for Sprogteknologi.
Kim, Y. (2009) Korean lexical bundles in conversation and academic text. *Corpora*, 4 (2): 135–65.

King, B. (2009) Building and analyzing corpora of computer-mediated communication. In P. Baker (ed.), *Contemporary Corpus Linguistics*, pp. 301–20. London: Continuum.
Kirk, J. (2002) Teaching critical skills in corpus linguistics using the BNC. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 155–63. Amsterdam: Rodopi.
Kirkpatrick, A. and Xu, Z. (2002) Chinese pragmatic norms and 'China English'. *World Englishes*, 21 (2): 269–79.
Kjellmer, G. (2005) Collocations and semantic prosody. Paper presented at the American Association of Applied Corpus Linguistics Conference, University of Michigan, Ann Arbor, 12–15 May.
Kniffka, H. with S. Blackwell and M. Coulthard (eds) (1997) *Recent Developments in Forensic Linguistics*. Frankfurt am Main: Peter Lang.
Koester, A. (2006) *Investigating Workplace Discourse*. London: Routledge.
Koester, A. (2007) 'About twelve thousand or so': vagueness in North American and UK offices. In J. Cutting (ed.), *Vague Language Explored*, pp. 40–61. Basingstoke: Palgrave Macmillan.
Koester, A. (2010) *Workplace Discourse*. London: Continuum.
Kredens, K. (2003) Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In B. Lewandowska-Tomaszczyk (ed.), *PALC 2001: Practical Applications in Language Corpora*, pp. 405–45. Frankfurt am Main: Peter Lang.
Kress, G. (2001) *Multimodal Discourse: the Modes and Media of Contemporary Communication*. London: Arnold.
Kretzschmar, W., Anderson, J., Beal, J., Corrigan, K., Opas-Hänninen, L. and Plichta, B. (2006) Collaboration on corpora for regional and social analysis. *Journal of English Linguistics*, 34 (3): 172–205.
Krishnamurthy, R. (1987) The process of compilation. In J. McH. Sinclair (ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing*, pp. 62–86. London: Collins ELT.
Krishnamurthy, R. (2002) The corpus revolution in EFL dictionaries. *Kernerman Dictionary News*, no. 10.
Krishnamurthy, R. and Kosem, I. (2007) Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes*, 6(4): 356–73.
Kübler, N. (2011a) Working with corpora for translation teaching in a French-speaking setting. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 62–80. London: Continuum.
Kübler, N. (ed.) (2011b) *Corpora, Language, Teaching and Resources: From Theory to Practice*. Bern: Peter Lang.
Kucěra, H. and Francis, W.N. (1967) *Computational Analysis of Present Day English*. Providence, RI: Brown University Press.
Lambrou, M. and Stockwell, P. (eds) (2007) *Contemporary Stylistics*. London: Continuum.
Langacker, R. (1986) An introduction to cognitive grammar. *Cognitive Science*, 10: 1–40.
Langacker, R. (1988) A usage-based model. In B. Rudzka-Ostyn (ed.), *Topics in Cognitive Linguistics*, pp. 127–61. Current Issues in Linguistic Theory 50. Amsterdam: Benjamins.
Langacker, R. (2000) A dynamic usage-based model. In M. Barlow and S. Kemmer (eds), *Usage-Based Models of Language*, pp. 1–64. Stanford: CLSI Publications.
Larsen-Freeman, D. and Cameron, L. (2008) *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.
Laviosa, S. (1997) How comparable can 'comparable corpora' be? *Target*, 9 (2): 289–319.
Laviosa, S. (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

Lawson, A. (2000) 'Die schöne Geschichte': a corpus-based analysis of Thomas Mann's *Joseph und seine Brüder*. In B. Dodd (ed.), *Working with German Corpora*, pp. 161–80. Birmingham: The University of Birmingham Press.

Lawson, A. (2001) Rethinking French grammar for pedagogy: the contribution of spoken corpora. In R. Simpson and J. Swales (eds), *Corpus Linguistics in North America*, pp. 179–94. Ann Arbor, Mich.: University of Michigan Press.

Lee, D. (2000) Modelling variation in spoken and written language: the Multi-Dimensional Approach revisited. Unpublished PhD thesis, University of Lancaster, UK.

Lee, D. (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.

Lee, D. (2008) Corpora and discourse analysis: new ways of doing old things. In V.K. Bhatia et al. (eds), *Advances in Discourse Studies*, pp. 86–99. London: Routledge.

Lee, D. and Chen, S. (2009) Making a bigger deal of the smaller words: function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, 18 (4): 281–96.

Lee, D. and Swales, J. (2006) A corpus-based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25 (1): 56–75.

Lee-Wong, S. (2005) Raising awareness of hedging in Chinese business letters: linguistic symbol of precision or politeness? In F. Bargiela-Chiappini and M. Gotti (eds), *Asian Business Discourse(s)*, pp. 271–90. Bern: Peter Lang.

Leech, G. (1991) The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics*, pp. 8–29. London: Longman.

Leech, G. (1992) Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 105–22. Berlin: Mouton de Gruyter.

Leech, G. (1997) Teaching and language corpora: a convergence. In A. Wichmann et al. (eds), *Teaching and Language Corpora*, pp. 1–23. London: Longman.

Leech, G. (2000) Grammar of spoken English: new outcomes of corpus-oriented research. *Language Learning*, 50 (4): 675–724.

Leech, G. and Short, M. (1981) *Style in Fiction*. London: Longman.

Leech, G. and Svartvik, J. (1994) *A Communicative Grammar of English*, 2nd edn. London: Longman.

Leistyna, P. and Meyer, C. (eds) (2003) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi.

Leńko-Szymańska, A. (2002) How to trace the growth in learners' active vocabulary: a corpus-based study. In B. Kettemann and G. Marko (eds) *Teaching and Learning by Doing Corpus Analysis*, pp. 217–30. Amsterdam: Rodopi.

Leńko-Szymańska, A. (2004) Demonstratives as anaphora markers in advanced learners' English. In G. Aston et al. (eds) *Corpora and Language Learners*, pp. 89–108. Amsterdam: John Benjamins.

Li, D. (2000) Hong Kong English: new variety of English or interlanguage? *English Australia Journal*, 18 (1): 50–9.

Lin, C.-Y. (2010) '…that's actually sort of you know trying to get consultants in…': Functions and multifunctionality of modifiers in academic lectures. *Journal of Pragmatics*, 42: 1173–83.

Lindemann, S. and Mauranen, A. (2001) 'It's just real messy': the occurrence and function of *just* in a corpus of academic speech. *English for Specific Purposes*, 20 (1): 459–75.

Lindquist, H. (2009) *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Liu, D. and Jiang, P. (2009) Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *The Modern Language Journal*, 93 (i): 61–78.

Longman Dictionary of Contemporary English (1995), 3rd edn. Harlow: Pearson.

Longman Essential Activator (1997) Longman: London.

Lopez-Ferrero, C. (2007) Academic writing: exploring *Corpus 92*. In G. Parodi (ed.), *Working with Spanish Corpora*, pp. 173–94. London: Continuum.

Lorenz, G. (1998) Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In S. Granger (ed.), *Learner English on Computer*, pp. 53–66. London: Longman.

Louw, B. (1993) Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker et al. (eds), *Text and Technology*, pp. 157–76. Amsterdam: John Benjamins.

Louw, B. (1997) The role of corpora in critical literary appreciation. In A. Wichmann et al. (eds), *Teaching and Language Corpora*, pp. 240–51. London: Longman.

Louw, B. (2006) *Corpus Approaches to the Language of Literature. Collocation, Corpora and Criticism*. AHDS Literature, Languages and Literature. Oxford Text Archive.

Lu, X. (2010) What can corpus software reveal about language development? In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 184–93. London: Routledge.

McCarthy, M. (1998) *Spoken Languages and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. (2000) Mutually captive audiences: small talk and the genre of close-contact service encounters. In J. Coupland (ed.), *Small Talk*, pp. 84–109. London: Longman Pearson.

McCarthy, M. (2001) *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. (2008) Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, 41(4): 563–74.

McCarthy, M. and Carter, R. (2002) 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.

McCarthy, M. and Carter, R. (2004) 'There's millions of them': hyperbole in everyday conversation. *Journal of Pragmatics*, 36: 149–84.

McCarthy, M. and Handford, M. (2004) 'Invisible to us': a preliminary corpus-based study of spoken business English. In U. Connor and T. Upton (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 167–201. Amsterdam: John Benjamins.

McCarthy, M., McCarten, J. and Sandiford, H. (2005) *Touchstone*. Cambridge: Cambridge University Press.

McCrostie, J. (2008) Writer visibility in EFL learner academic writing: a corpus-based study. *ICAME Journal*, 32: 97–114.

McEnery, T. (2000) Public lecture on corpus linguistics. Chinese University of Hong Kong, 18 April.

McEnery, T. (2007) Press talk: the UK Press and refugees. Public lecture delivered at the Department of English and Communication, City University of Hong Kong, 26 February 2007.

McEnery, T. and Kifle, N. (2002) Epistemic modality in argumentative essays of second-language writers. In J. Flowerdew (ed.), *Academic Discourse*, pp. 182–95. London: Longman.

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): 403–19.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics*, 2nd edn. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y (2006) *Corpus-Based Language Studies*. London: Routledge.

McKenny, J. (2003) Swift's Prescience: a polite precursor of corpus linguistics. *Journal of Language and Literature*, 2 (1): 1–11.

McKenny, J. and Bennett, K. (2011) Polishing papers for publication: palimpsests or procrustean beds? In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 247–62. London: Continuum.

Mackenzie, I. (2003) English as a lingua franca. *The European Messenger*, 12 (1): 59–61.

Mahlberg, M. (2007a) Corpus stylistics: bridging the gap between linguistic and literary studies. In M. Hoey et al. (eds), *Text, Discourse and Corpora*, pp. 219–46. London: Continuum.

Mahlberg, M. (2007b) Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2 (1): 1–31.

Mahlberg, M. (2010) A corpus stylistic perspective on Dickens's *Great Expectations*. In M. Lambrou and P. Stockwell (eds), *Contemporary Stylistics*, 19–38. London: Continuum.

Maier, P. (1992) Politeness strategies in business letters by native and non-native English speakers. *English for Specific Purposes*, 11: 189–205.

Mair, C. (2002) Empowering non-native speakers: the hidden surplus value of corpora in continental English departments. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 119–30. Amsterdam: Rodopi.

Mair, C. (2006) Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. In A. Renouf and A. Kehoe (eds), *The Changing Face of Corpus Linguistics*, pp. 355–76. Amsterdam: Rodopi.

Mair, C. (2009) Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change. In A. Renouf and A. Kehoe (eds), *Corpus Linguistics. Refinements and Reassessments*, pp. 7–32. Amsterdam: Rodopi.

Mair, C. and Hundt, M. (eds) (2000) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Maley, A. (2009) ELF: a teacher's perspective. *Language and Intercultural Communication*, 9 (3): 187–200.

Malmkjær, K. (1998) Love thy neighbour: will parallel corpora endear linguists to translators? *Meta*, 43 (4): 534–41.

Malmkjær, K. (2009) On a pseudo-subversive use of corpora in translator training. In F. Zanettin et al. (eds), *Corpora in Translator Education*, pp. 119–34. Manchester, UK: St Jerome Publishing.

Marra, M. and Holmes, J. (2002) Laughing on the inside: humour and internal politics of the workplace. Language in the Workplace Occasional Papers 4. Victoria University of Wellington.

Martin, J. and White, P. (2005). *The Language of Evaluation. Appraisal in English*. Basingstoke: Palgrave Macmillan.

Matthiessen, C. (1996) The relationship between grammar and discourse: an exploration based on computational tools. Keynote lecture given at the University of Macau, 15 June.

Mauranen, A. (2001) Reflexive academic talk: observations from MICASE. In R. Simpson and J. Swales (eds), *Corpus Linguistics in North America*, pp. 165–78. Ann Arbor, Mich.: University of Michigan Press.

Mauranen, A. (2003a) 'But here's a flawed argument': socialisation into and through metadiscourse. In P. Leistyna and C. Meyer (eds), *Corpus Analysis: Language Structure and Language Use*, pp. 19–34. Amsterdam: Rodopi.

Mauranen, A. (2003b) The corpus of English as lingua franca in academic settings. *TESOL Quarterly*, 37 (3): 513–27.

Mauranen, A. (2004a) Spoken corpus for an ordinary learner. In J. Sinclair (ed.), *How to Use Corpora in Language Teaching*, pp. 89–105. Amsterdam: John Benjamins.

Mauranen, A. (2004b) 'They're a little bit different'…: Observations on hedges in academic talk. In K. Aijmer and A–B. Stenström (eds), *Discourse Patterns in Spoken and Written Corpora*, pp. 173–97. Amsterdam: John Benjamins.

Mauranen, A. (2007) Investigating English as a lingua franca with a spoken corpus. In M.C. Campoy and M.J. Luzón (eds), *Spoken Corpora in Applied Linguistics*, pp. 33–56. Berlin: Peter Lang.

Mauranen, A., Hynninen, N. and Ranta, E. (2010) English as an academic lingua franca: the ELFA project. *English for Specific Purposes*, 29 (3): 183–90.

Mautner, G. (2009a) Corpora and critical discourse analysis. In P. Baker (ed.), *Contemporary Corpus Linguistics*, pp. 32–46. London: Continuum.

Mautner, G. (2009b) Checks and balances: how corpus linguistics can contribute to CDA. In R. Wodak and M. Meyer (eds), *Methods of Critical Discourse Analysis*, pp. 122–43. London: Sage.

Meunier, F. and Gouverneur, C. (2009) New types of corpora for new educational challenges. Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (ed.), *Corpora and Language Teaching*, pp. 179–21. Amsterdam: John Benjamins.

Meunier, F. and Granger, S. (eds) (2008) *Phraseology in Foreign Language Teaching and Learning*. Amsterdam: John Benjamins.

Meyer, C. (2002) *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, G. and Chomsky, N. (1963) Finitary models of language users. In R. Luce, R. Bush and E. Galanter (eds) *Handbook of Mathematical Psychology*, Vol. ii, pp. 419–91. New York: Wiley.

Milton, J. (2004) From parrots to puppet masters: fostering creative and authentic language use with online tools. In B. Homberg, M. Shelley and C. White (eds), *Distance Education and Languages: Evolution and Change*, pp. 242–57. Clevedon: Multilingual Matters.

Milton, J. and Hyland, K. (1999) Assertions in students' academic essays: a comparison of English NS and NNS student writers. In R. Berry, B. Asker and K. Hyland (eds) *Language Analysis, Description and Pedagogy*, pp. 147–61. Hong Kong: Language Centre, HKUST.

Mishan, F. (2004) Authenticating corpora for language learning: a problem and its solution. *ELT Journal*, 58 (3): 219–27.

Mollin, S. (2006) *Euro-English: Assessing Variety Status*. Gunter Narr Verlag.

Mollin, S. (2007) English as a lingua franca: a new variety in the new expanding circle? *Nordic Journal of English Studies*, 5 (2): 41–57.

Moon, R. (1987) The analysis of meaning. In J. McH. Sinclair (ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing*, pp. 86–103. London: Collins ELT.

Moon, R. (1998) *Fixed Expressions and Idioms in English: a Corpus-Based Approach*. Oxford Studies in Lexicography and Lexicology. Oxford: Oxford University Press.

Morley, B. (2006) WebCorp: a tool for online linguistic information retrieval and analysis. In A. Renouf and A. Kehoe (eds), *The Changing Face of Corpus Linguistics*, pp. 283–96. Amsterdam: Rodopi.

Morley, J. and Bailey, P. (eds) (2009) *Corpus-Assisted Discourse Studies on the Iraq Conflict*. London: Routledge.

Moss, L. (2009) Henry James beyond the numbers: applying corpus analysis to the text. Paper presented in colloquium 'Corpus Linguistics and Literature'. Corpus Linguistics Conference, University of Birmingham, UK, 20 July.

Mudraya, O. (2006) Engineering English: a lexical frequency instructional model. *English for Specific Purposes*, 25 (2): 23–56.

Mukherjee, J. (2004a) Corpus data in a usage-based cognitive grammar. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 85–100. Amsterdam: Rodopi.

Mukherjee, J. (2004b) Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor and T. Upton (eds), *Applied Corpus Linguistics: a Multidimensional Perspective*, pp. 239–50. Amsterdam: Rodopi.

Mukherjee, J. (2005) The native speaker is alive and kicking – linguistic and language pedagogic perspective. *Anglistik*, 16 (2): 7–23.

Mukherjee, J. (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.

Mukherjee, J. (2009) The grammar of conversation in advanced spoken learner English. Learner corpus data and language-pedagogical implications. In K. Aijmer (ed.), *Corpora and Language Teaching*, pp. 203–30. Amsterdam: John Benjamins.

Mukherjee, J. and Rohrback, J. (2006) Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research. In B. Kettemann and G. Marko (eds), *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*, pp. 205–31. Frankfurt am Main: Peter Lang.

Mur Dueñas, P. (2009) Logical markers in L1 (Spanish and English) and L2 (English) business research articles. *English Text Construction*, 2 (2): 246–64.

Myles, F. (2005) Interlanguage corpora and second language acquisition research. Review article. *Second Language Research*, 21 (4): 373–91.

Myles, F. and Mitchell, R. (2004) Using information technology to support empirical SLA research. *Journal of Applied Linguistics*, 1 (2): 169–96.

Nattinger, J. (1980) A lexical phrase grammar for ESL. *TESOL Quarterly*, 14: 337–44.

Nelson, M. (2006) Semantic associations in business English: a corpus-based analysis. *English for Specific Purposes*, 25 (2): 217–34.

Nesi, H. (2011) BAWE: an introduction to a new resource. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 213–28. London: Continuum.

Nesselhauf, N. (2003) The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24 (2): 223–42.

Nesselhauf, N. (2004) Learner corpora and their potential for language teaching. In J. McH. Sinclair (ed.), *How to Use Corpora in Language Teaching*, pp. 125–52. Amsterdam: John Benjamins.

Nesselhauf, N. (2005) *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

New Oxford Dictionary of English (1998) Oxford: Oxford University Press.

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds), pp. 572–90. UCREL Technical Paper number 16, Lancaster University.

Noguchi, J. (2004) A genre-analysis and mini-corpora approach to support professional writing by nonnative English speakers. *English Corpus Studies*, 11: 1–11.

Noguchi, J., Orr, T. and Tono, Y (2006) Using a dedicated corpus to identify features of professional English usage. In A. Wilson, D. Archer and P. Rayson (eds), *Corpus Linguistics around the World*, pp. 155–66. Amsterdam: Rodopi.

Norris, S. (2004) *Analysing Multimodal Interaction: a Methodological Framework*. London: Routledge.

Oakey, D. (2009) Fixed collocational patterns in isolexical and isotextual versions of a corpus. In P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 140–58. London: Continuum.

OALD (1974) *Oxford Advanced Learner's Dictionary of Current English*, 3rd edn. Oxford: Oxford University Press.

OALD (2000) *Oxford Advanced Learner's Dictionary of Current English*, 6th edn. Oxford: Oxford University Press.

O'Halloran, K. (2007) The subconscious in James Joyce's 'Eveline': a corpus stylistics analysis that chews on the 'Fish Hook'. *Language and Literature*, 16 (3): 227–44.

O'Halloran, K. (2009) Inferencing and cultural reproduction: a corpus-based critical discourse analysis. *Text and Talk*, 29 (1): 21–51.

O'Halloran, K. and Coffin, C. (2004) Checking overinterpretation and underinterpretation: help from corpora in critical linguistics. In C. Coffin et al. (eds) *Applying English Grammar: Functional and Corpus Approaches*, pp. 275–97. The Open University: Arnold.

O'Keeffe, A. (2006) *Investigating Media Discourse*. London: Routledge.

O'Keeffe, A. and Farr, F. (2003) Using language corpora in initial teacher education: pedagogic issues and practical applications. *TESOL Quarterly*, 37 (3): 389–418.

O'Keeffe, A. and McCarthy, M. (eds) (2010) *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Olohan, M. (2003) Spelling out the optionals in translation: a corpus study. In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds), *Proceedings of the Corpus Linguistics 2001 Conference*, pp. 423–32. UCREL: Lancaster University.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge.

Olohan, M. and Baker, M. (2000) Reporting *that* in translated English: evidence for subconscious processes of explicitation? *Across Languages and Cultures*, 1 (2): 141–58.

Ooi, V. (1997) Analysing the Singapore ICE corpus for lexicographic evidence. In M. Ljung (ed.), *Corpus-Based Studies in English*, pp. 245–59. Amsterdam: Rodopi.

Ooi, V. (2009) Computer-mediated language and corpus linguistics. In Y. Kawaguchi, M. Minegishi and J. Durand (eds), *Corpus Analysis and Variation in Linguistics*. Amsterdam: John Benjamins, pp. 103–20.

Ooi, V., Tan, P. and Chiang, A. (2007) Analysing personal weblogs in Singapore English: the Wmatrix approach. In *eVariEng* (*Journal of the Research Unit for Variation, Contacts, and Change in English*), vol. 2, *Towards Multimedia in Corpus Studies*. Finland: University of Helsinki.

Orton, H. (1962) *Survey of English Dialects: an Introduction*. Leeds: E. J. Arnold.

Osborne, J. (2000) What can students learn from a corpus?: building bridges between data and explanation. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 165–72. Frankfurt am Main: Peter Lang.

Osborne, J. (2004) Top-down and bottom-up approaches to corpora in language teaching. In U. Connor and T. Upton (eds), *Applied Corpus Linguistics: a Multidimensional Perspective*, pp. 251–65. Amsterdam: Rodopi.

Osborne, J. (2011a) Oral learner corpora and assessment of speaking skills. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 181–97. London: Continuum.

Osborne, J. (2011b) Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics*, 16 (2): 276–98.

O'Sullivan, I. and Chambers, A. (2006) Learners' writing skills in French: corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15 (1): 49–68.

Owen, C. (1996) Does a corpus require to be consulted? *ELT Journal*, 50: 219–24.
Oxford Collocations Dictionary (2002) Oxford: Oxford University Press.
Oxford English Dictionary (1989). Oxford: Oxford University Press.
Oxford–Hachette French Dictionary (1994) Oxford: Oxford University Press.
Paltridge, B. (2006) *Discourse Analysis*. London: Continuum.
Paquot, M. (2010) *Academic Vocabulary in Learner Writing*. London: Continuum.
Parodi, G. (2007) Variation across registers in Spanish. Exploring the El Grial PUCV Corpus. In G. Parodi (ed.), *Working with Spanish Corpora*, pp. 11–53. London: Continuum.
Parodi, G. (ed.) (2010) *Academic and Professional Discourse Genres in Spanish*. Amsterdam: John Benjamins.
Partington, A. (1998) *Patterns and Meanings*. Amsterdam: John Benjamins.
Partington, A. (2003) *The Linguistics of Political Argument. The Spin-Doctor and the Wolf-Pack at the White House*. London and New York: Routledge.
Partington, A. (2004a) Utterly content in each other's company. *International Journal of Corpus Linguistics*, 9 (1): 131–56.
Partington, A. (2004b) Corpora and discourse, a most congruous beast. In A. Partington, J. Morley and L. Haarman (eds), *Corpora and Discourse*, pp. 11–20. Bern: Peter Lang.
Pawley, A. and Syder, F.H. (1983) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds), *Language and Communication*, pp. 191–226. London: Longman.
Pearce, M. (2008) Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine. *Corpora*, 3 (1): 1–29.
Pearson, J. (1998) *Terms in Context*. Amsterdam: John Benjamins.
Pearson, J. (2000) Teaching terminology using electronic resources. In S. Botley, T. McEnery and A. Wilson (eds) *Multilingual Corpora in Teaching and Research*, pp. 92–105. Amsterdam: Rodopi.
Pérez Basanta, C. and Rodriguez Martin, M. (2007) the application of data-driven learning to a small–scale corpus: using film transcripts for teaching conversational skills. In E. Hidalgo et al. (eds), *Corpora in the Foreign Language Classroom*, pp. 141–58. Amsterdam: Rodopi.
Pérez-Paredes, P., Sánchez-Tornel, M., Alcarez Colero, J. and Aguado Jiménez, P. (2011) Tracking learners' actual uses of corpora: guided vs. non-guided corpus consultation. *Computer-Assisted Language Learning,* 24(3): 233–53.
Peters, P. (1996) Comparative insights into comparison. *World Englishes*, 15 (1): 57–67.
Pinker, S. (1999) *Words and Rules: the Ingredients of Language*. Weidenfeld and Nicolson.
Plevoets, K, Speelman, D. and Geeraerts, D. (2008) The distribution of T/V pronouns in Netherlandic and Belgian Dutch. In K. Schneider and A. Barron (eds), *Variational Pragmatics*, pp. 181–209. Amsterdam: John Benjamins.
Poncini, G. (2004) *Discursive Strategies in Multicultural Business Meetings*. Bern: Peter Lang.
Poos, D. and Simpson, R. (2002) Cross-disciplinary comparisons of hedging: some findings from the Michigan Corpus of Academic Spoken English. In R. Reppen et al. (eds), *Using Corpora to Explore Linguistic Variation*, pp. 3–23. Amsterdam: John Benjamins.
Popper, K. (1963) *Conjectures and Refutations: the Growth of Scientific Knowledge*. London: Routledge and Kegan Paul.
Pravec, N. (2002) Survey of learner corpora. *ICAME Journal*, 26: 8–14.
Prodromou, L. (2008) *English as a Lingua Franca: a Corpus-Based Analysis*. London: Continuum.
Quirk, R (1992) On corpus principles and design. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 457–69. Berlin: Mouton de Gruyter.
Quirk, R., Greenbaum, S., Leech G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Rayson, P. (2005) WMatrix. A web-based corpus processing environment. Computing Department, Lancaster University. http://ucrel/lancs.ac.uk/wmatrix/

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. *International Journal of Corpus Linguistics*, 2: 133–50.

Reinhardt, J. (2010) Directives in office hour consultations: a corpus-informed investigation of learner and expert usage. *English for Specific Purposes*, 29 (2): 94–107.

Renouf, A. (1997) Teaching corpus linguistics to teachers of English. In A. Wichmann et al. (eds), *Teaching and Language Corpora*, pp. 255–66. London: Longman.

Renouf, A. (2003) WebCorp: providing a renewable data source for corpus linguists. *Language and Computers*, 48: 39–58.

Renouf, A. and Kehoe, A. (eds) (2006) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.

Renouf, A. and Kehoe, A. (eds) (2009) *Corpus Linguistics. Refinements and Reassessments*. Amsterdam: Rodopi.

Renouf, A., Kehoe, A. and J. Banerjee (2007) WebCorp: an integrated system for web text search. In M. Hundt et al. (eds), *Corpus Linguistics and the Web*, pp. 47–67. Amsterdam: Rodopi.

Renouf, A. and Sinclair, J. McH. (1991) Collocational frameworks in English. In K. Aijmer and B. Altenberg (eds), *Advances in Corpus Linguistics*, pp. 128–43. Amsterdam: Rodopi.

Reppen, R. (2001) Writing development among elementary students: corpus-based perspectives. In R. Simpson and J. Swales (eds), *Corpus Linguistics in North America*, pp. 211–25. Ann Arbor, Mich.: University of Michigan Press.

Reppen, R. (2010) *Using Corpora in the Language Classroom*. New York: Cambridge University Press.

Reppen, R., Fitzpatrick, S. and Biber, D. (eds) (2002) *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Reppen, R. and Ide, N. (2004) American National Corpus: overall goals and the first release. *Journal of English Linguistics*, 32(2): 105–13.

Rissanen, M. (2009) Corpus linguistics and historical linguistics. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook*, vol. 1, pp. 53–68. Berlin: Mouton de Gruyter.

Rodríguez, P. (2006) The application of electronic corpora to translation teaching within a task-based approach. In A. Hornero, M. Luzón and S. Murillo (eds), *Corpus Linguistics: Applications for the Study of English*, pp. 301–12. Bern: Peter Lang.

Römer, U. (2004) A corpus-driven approach to modal auxiliaries and their didactics. In J. McH. Sinclair (ed.), *How to Use Corpora in Language Teaching*, pp. 185–99. Amsterdam: John Benjamins.

Römer, U. (2006) Pedagogical applications of corpora: some reflections on the current scope and a wish list for future developments. *Zeitschrift Anglistik und Americanstik*, 54 (2): 121–34.

Römer, U. (2010) Using general and specialized corpora in English language teaching: past, present and future. In M. Campoy-Cubillo et al. (eds), *Corpus-Based Approaches to English Language Teaching*, pp. 18–38. London: Continuum.

Römer, U. and Schulze, R. (eds) (2009) *Exploring the Lexis–Grammar Interface*. Amsterdam: John Benjamins.

Rose, D. (2008) Vocabulary use in the FCE listening test. *Research Notes*, 32: 9–16. Accessed 3 July, http://www.cambridgeesol.org/rs_notes/

Rühlemann, C. (2007) *Conversation in Context: a Corpus-Driven Approach*. Amsterdam: John Benjamins.

Rundell, M. (ed.) (2002) *Macmillan English Dictionary*. London: Macmillan.
Salamoura, A. (2008) Aligning English Profile research data to the CEFR. *Research Notes*, 33: 5–7. Accessed 3 July, http://www.cambridgeesol.org/rs_notes/
Sampson, G. (1996) From central embedding to corpus linguistics. In J. Thomas and M. Short (eds), *Using Corpora for Language Research*, pp. 14–26. London: Longman.
Sampson, G. (2001) *Empirical Linguistics*. London and New York: Continuum.
Sampson, G. (2005) *The 'Language Instinct' Debate* (revised edn). London: Continuum.
Sampson, G. and McCarthy, D. (eds) (2004) *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum.
Santos, D. and Frankenberg-Garcia, A. (2007) The corpus, its users and their needs: a user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12 (3): 335–74.
Sarangi, S. and Roberts, C. (eds) (1999) *Talk Work and Institutional Order: Discourse in Medical, Mediation and Management Settings*. Berlin: Mouton de Gruyter.
Schmid, H. (2000) *English Abstract Nouns as Conceptual Shells: from Corpus to Cognition*. Berlin: Mouton de Gruyter.
Schmied, J. (1990) Corpus linguistics and non-native varieties of English. *World Englishes*, 9 (3): 255–68.
Schmied, J. (2006a) Corpus linguistics and grammar learning: tutor versus learner perspectives. In S. Braun et al. (eds), *Corpus Technology and Language Pedagogy*, pp. 87–106. Frankfurt am Main: Peter Lang.
Schmied, J. (2006b) New ways of analyzing ESL on the WWW with WebCorp and WebPhraseCount. In A. Renouf and A. Kehoe (eds), *The Changing Face of Corpus Linguistics*, pp. 309–24. Amsterdam: Rodopi.
Schmied, J. and Schäffler, H. (1996) Approaching translationese through parallel and translation corpora. In C. Percy, C. Meyer and I. Lancashire (eds), *Synchronic Corpus Linguistics*, pp. 41–56. Amsterdam: Rodopi.
Schmitt, N. (ed.) (2004) *Formulaic Sequences*: *Acquisition, Processing and Use*. Amsterdam: John Benjamins.
Schmitt, N. (2010) *Researching Vocabulary: a Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.
Schmitt, N., Grandage, S. and Adolphs, S. (2004) Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (ed.), *Formulaic Sequences*: *Acquisition, Processing and Use*, pp. 127–51. Amsterdam: John Benjamins.
Schneider, K. and Barron, A. (eds) (2008) *Variational Pragmatics*. Amsterdam: John Benjamins.
Schönefeld, D. (1999) Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics*, 4 (1): 137–71.
Scott, M. (1997) PC analysis of key words – and key key words. *System*, 25 (1): 1–13.
Scott, M. (1999) *WordSmith Tools* (version 3.0). Oxford: Oxford University Press.
Scott, M. (2001) Comparing corpora and identifying key words, collocations and, frequency distributions through the *WordSmith Tools* suite of computer programs. In M. Ghadessy et al. (eds), *Small Corpus Studies and ELT*, pp. 47–67. Amsterdam: John Benjamins.
Scott, M. (2004) *WordSmith Tools* (version 4.0). Oxford: Oxford University Press.
Scott, M. and Tribble, C. (2006) *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Sealey, A. (2009) Probabilities and surprises: a realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics*, 31 (2): 215–35.

Sealey, A. and Thompson, P. (2004) 'What do you call the dull words?' Primary school children using corpus-based approaches to learn about language. *English in Education*, 38 (1): 80–91.

Sealey, A. and Thompson, P. (2007) Corpus, concordance, classification: young learners in the L1 classroom. *Language Awareness*, 16 (3): 208–23.

Seidlhofer, B. (2000) Operationalising intertextuality: using learner corpora for learning. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 207–23. Frankfurt am Main: Peter Lang.

Seidlhofer, B. (2001a) Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11: 133–58.

Seidlhofer, B. (2001b) Making the case for a corpus of English as a lingua franca. In G. Aston and L. Burnard (eds), *Corpora in the Description and Teaching of English*, pp. 70–85. Cooperativa Libraria Universitaria Editrice Bologna.

Seidlhofer, B. (2002) Pedagogy and local learner corpora: working with learning-driven data. In S. Granger et al. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, pp. 213–34. Amsterdam: John Benjamins.

Seidlhofer, B. (ed.) (2003) *Controversies in Applied Linguistics* (Ch. 2: Corpus linguistics and language teaching). Oxford: Oxford University Press.

Seidlhofer, B. (2004) Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics*, 24: 209–39.

Seidlhofer, B. (2010) Orientations in ELF research: form and function. In A. Mauranen and E. Ranta (eds) *English as a Lingua Franca*, pp. 37–59. Newcastle upon Tyne: Cambridge Scholars Publishing.

Seidlhofer, B. and Jenkins, J. (2003) English as a lingua franca and the politics of property. In C. Mair (ed.), *The Politics of English as a World Language*, pp. 139–54. Amsterdam: Rodopi.

Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Narratives*. London: Routledge.

Shortall, T. (2005) Corpus, curriculum, and the language instinct. Paper presented at the Corpus Linguistics 2005 Conference. University of Birmingham, UK.

Shortall, T. (2007) The L2 syllabus: corpus or contrivance? *Corpora*, 2: 157–85.

Shulman, L. (1987) Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, 57 (1): 1–22.

Sifakis, N. (2007) The education of teachers of English as a lingua franca: a transformative perspective. *International Journal of Applied Linguistics*, 17 (3): 355–75.

Simpson, R. and Swales, J. (eds) (2001) *Corpus Linguistics in North America*. Ann Arbor, Mich.: University of Michigan Press.

Simpson-Vlach, R. and Ellis, N. (2010) An academic formulas list: new methods in phraseological research. *Applied Linguistics*, 31 (4): 487–512.

Sinclair, J. McH. (1985) Selected issues. In R. Quirk and H.G. Widdowson (eds), *English in the World*. Cambridge: Cambridge University Press.

Sinclair, J. McH. (ed.) (1987) *Looking Up. An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Sinclair, J. McH. (ed.-in-chief) (1990) *Collins COBUILD English Grammar*. London: HarperCollins.

Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. McH. (1992) The automatic analysis of corpora. In J. Svartvik (ed.), *Directions in Corpus Linguistics*, pp. 379–97. Berlin: Mouton de Gruyter.

Sinclair, J. McH. (1998) Large corpus research and foreign language teaching. In R. de Beaugrande, M. Grosman, and B. Seidlhofer (eds), *Language Policy and Language Education in Emerging* Nations, pp. 79–86. London: Ablex.

Sinclair, J. McH. (1999) The lexical item. In E. Weigand (ed.), *Contrastive Lexical Semantics*, pp. 1–24. Amsterdam: John Benjamins.

Sinclair, J. McH. (2002) Review of *The Longman Grammar of Spoken and Written English*. *International Journal of Corpus Linguistics*, 6 (2): 339–59.

Sinclair, J. McH. (2003a) Corpora for lexicography. In P. van Sterkenberg (ed.), *A Practical Guide to Lexicography*, pp. 167–78. Amsterdam: John Benjamins.

Sinclair, J. McH. (2003b) Corpus processing. In P. van Sterkenberg (ed.), *A Practical Guide to Lexicography*, pp. 179–93. Amsterdam: John Benjamins.

Sinclair, J. McH. (2004a) *Trust the Text*. London and New York: Routledge.

Sinclair, J. McH. (ed.) (2004b) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Sinclair, J. McH. (2005) Corpus and Text – Basic Principles. In M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pp. 1–21. Oxford Text Archive. Available at: http://ahds.ac.uk/linguistic-corpora/.

Sinclair, J. McH. and Coulthard, M. (1975) *Towards an Analysis of Discourse*. Oxford: Oxford University Press.

Sinclair, J. McH., Hoey, M. and Fox, G. (eds) (1993) *Techniques of Description – Spoken and Written Discourse*. London: Routledge.

Sinclair, J. McH. and Renouf, A. (1988) A lexical syllabus for language learning. In R. Carter and M. McCarthy (eds), *Vocabulary and Language Teaching*, pp. 140–60. London: Longman.

Skelton, J. and Hobbs, F. (1999) Concordancing: use of language-based research in medical communication. *The Lancet*, vol. 353: 108–11.

Skelton, J., Wearn, A. and Hobbs, R. (2002) 'I' and 'we': a concordancing analysis of how doctors and patients use first person pronouns in primary care consultations. *Family Practice*, 19 (5): 484–8.

Solan, L. and Tiersma, M. (2004) Author identification in American courts. *Applied Linguistics*, 25 (4): 448–65.

Spencer-Oatey, H. and Franklin, P. (2009) *Intercultural Interaction: a Multidisciplinary Approach to Intercultural Communication*. Basingstoke: Palgrave Macmillan.

Sripicharn, P. (2010) How can we prepare learners for using language corpora? In A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, pp. 371–84. London: Routledge.

Starfield, S. (2004) Why does this feel empowering? Thesis writing, concordancing and the 'corporatising university'. In B. Norton and K Toohey (eds), *Critical Pedagogies and Language Learning*, pp. 138–57. Cambridge: Cambridge University Press.

Starcke, B. (2006) The phraseology of Jane Austen's *Persuasion*: phraseological units as carriers of meaning. *ICAME Journal*, 3087–104.

Stenström, A.-B. (1994) *An Introduction to Spoken Interaction*. London: Longman.

Stewart, D. (2009) *Semantic Prosody. A Critical Evaluation*. London: Routledge.

Strecker, B. (1985) Rules and the dynamics of language. In T. Ballmer (ed.), *Linguistic Dynamics*, pp. 238–50. New York: Mouton de Gruyter.

Stubbe, M. (2001) From office to production line: collecting data for the Wellington Language in the Workplace Project. Language in the Workplace Occasional Papers 2. Victoria University of Wellington.

Stubbs, M. (1983) *Discourse Analysis*. Chicago: University of Chicago Press.

Stubbs, M. (1995a) Corpus evidence for norms of lexical collocation. In G. Cook and B. Seidlhofer (eds), *Principle and Practice in Applied Linguistics*, pp. 245–56. Oxford: Oxford University Press.

Stubbs, M. (1995b) Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2 (1): 23–55.

Stubbs, M. (1996) *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. (2001a) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2001b) On inference theories and code theories: corpus evidence for semantic schemas. *Text*, 21 (3): 437–65.

Stubbs, M. (2004) Language corpora. In A. Davies and C. Elder (eds), *The Handbook of Applied Linguistics*, pp. 106–32. Malden, Mass.: Blackwell.

Stubbs, M. (2005) Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14 (1): 5–24.

Stubbs, M. (2007) Quantitative data on multi-word sequences in English: the case of the word *world*. In M. Hoey et al. (eds), *Text, Discourse and Corpora*, pp. 163–89. London: Continuum.

Stubbs, M. (2009) The search for units of meaning. Sinclair on empirical semantics. *Applied Linguistics*, 30 (1): 115–37.

Sung Park, E. (2004) The comparative fallacy in UG studies. Working Papers in TESOL and Applied Linguistics 4 (1). Available from: http://www.tc.columbia.edu/academic/tesol/Webjournal/forum2004.htm.

Svartvik, J. (ed.) (1992) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.

Swales, J. (1990) *Genre Analysis*. Cambridge: Cambridge University Press.

Swales, J. (2002) Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (ed.), *Academic Discourse*, pp. 150–64. London: Longman.

Swales, J. (2004) *Research Genres*. Cambridge: Cambridge University Press.

Szakos, J. (2000) Producing and using corpora in Chinese language education. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 187–92. Frankfurt am Main: Peter Lang.

Tagliamonte, S. and Lawrence, S. (2000) 'I Used to Dance, but I Don't Dance Now': the habitual past in English. *Journal of English Linguistics*, 28 (4): 324–53.

Takaie, H. (2002) A trap in corpus linguistics: the gap between corpus-based analysis and intuition-based analysis. In T. Saito et al. (eds), *English Corpus Linguistics in Japan*, pp. 111–30. Amsterdam: Rodopi.

Taylor, L. and Barker, F. (2008) Using corpora for language assessment. In E. Shohamy and N. Hornberger (eds), *Encyclopedia of Language and Education*, pp. 241–54. Boston, Mass.: Springer Science + Business Media LLC.

Teubert, W. (2004) Units of meaning, parallel corpora, and their implications for language teaching. In U. Connor and T. Upton (eds), *Applied Corpus Linguistics: a Multidimensional Perspective*, pp. 171–89. Amsterdam: Rodopi.

Teubert, W. (2005) My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10 (1): 1–13.

Thomas, J. (1983) Cross-cultural pragmatic failure. *Applied Linguistics*, 4: 91–112.

Thomas, J. and Boulton, A. (eds) (2012) *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.

Thomas, J. and Short, M. (eds) (1996) *Using Corpora for Language Research*. London: Longman.

Thomas, J. and Wilson, A. (1996) Methodologies for studying a corpus of doctor–patient interaction. In J. Thomas and M. Short (eds), *Using Corpora for Language Research*, pp. 92–109. London: Longman.

Thompson, G. and Hunston, S. (eds) (2006) *System and Corpus: Exploring Connections*. Equinox.

Thompson, P. and Sealey, A. (2007) Through children's eyes? Corpus evidence of the features of children's literature. *International Journal of Corpus Linguistics*, 12 (1): 1–23.

Thurstun, J. and Candlin, C. (1998a) *Exploring Academic English: a Workbook for Student Essay Writing*. Macquarie University: NCELTR.

Thurstun, J. and Candlin, C. (1998b) Concordancing and the teaching of vocabulary of academic English. *English for Specific Purposes*, 17 (3): 267–80.

Tognini Bonelli, E. (1993) Interpretative nodes in discourse – *Actual* and *Actually*. In M. Baker et al. (eds), *Text and Technology*, pp. 193–212. Amsterdam: John Benjamins.

Tognini Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tognini Bonelli, E. (2002) Functionally complete units of meaning across English and Italian: towards a corpus-driven approach. In B. Altenberg and S. Granger (eds), *Lexis in Contrast: Corpus-Based Approaches*, pp. 73–95. Amsterdam: John Benjamins.

Tognini Bonelli, E. (2004) Working with corpora: issues and insights. In C. Coffin et al. (eds), *Applying English Grammar: Functional and Corpus Approaches*, pp. 11–24. The Open University: Arnold.

Tognini Bonelli, E. and Sinclair, J. McH. (2006) Corpora. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn, pp. 216–19. Amsterdam: Elsevier.

Tono, Y. (2000) A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 123–32. Frankfurt am Main: Peter Lang.

Toolan, M. (2010) *Narrative Progression in the Short Story: a Corpus Stylistic Approach*. Amsterdam: John Benjamins.

Toury, G. (1980) *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Tracy-Ventura, N., Cortes, V. and Biber, D. (2007) Lexical bundles in Spanish speech and writing. In G. Parodi (ed.), *Working with Spanish Corpora*, pp. 354–75. London: Continuum.

Trebits, A. (2009a) Conjunctive cohesion in English language EU documents – a corpus-based analysis and its implications. *English for Specific Purposes*, 28: 199–210.

Trebits, A. (2009b) The most frequent phrasal verbs in English language EU documents. A corpus-based analysis and its implications. *System*, 37 (3): 470–81.

Tribble, C. (2000) Genres, keywords, teaching: towards a pedagogic account of language of project proposals. In L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective*, pp. 75–90. Frankfurt am Main: Peter Lang.

Tse, P. and Hyland, K. (2006) So what is the problem this book addresses? Interactions in academic book reviews. *Text and Talk*, 26 (6): 767–90.

Tsui, A. (2004) What teachers have also wanted to know – and how corpora can help. In J. McH. Sinclair (ed.), *How to Use Corpora in Language Teaching*, pp. 39–61. Amsterdam: John Benjamins.

Tsui, A. (2005) ESL teachers' questions and corpus evidence. *International Journal of Corpus Linguistics*, 10 (3): 335–56.

Tucker, G. (1996) So grammarians haven't the faintest idea: reconciling lexis-oriented and grammar-oriented approaches to language. In R. Hasan, C. Cloran and D. Butt (eds), *Functional Descriptions: Theory in Practice*, pp. 145–78. Amsterdam: John Benjamins.

Tucker, G. (2001) Possibly alternative modality. *Functions of Language*, 8 (2): 183–216.

Tyler, S.A. (1987) *The Unspeakable*. Madison: University of Wisconsin Press.

Upton, T. and Connor, U. (2001) Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20 (4): 313–29.

Utka, A. (2004) Phases of translation corpus. Compilation and analysis. *International Journal of Corpus Linguistics*, 9 (2): 195–224.

van Rij-Heyligers, J. (2007) To weep perilously or W.EAP critically: the case for a corpus-based critical EAP. In E. Hidalgo et al. (eds), *Corpora in the Foreign Language Classroom*, pp. 105–18. Amsterdam: Rodopi.

Van Sterkenberg, P. (ed.) (2003) *A Practical Guide to Lexicography*. Amsterdam: John Benjamins.

Varantola, K. (2003) Translators and disposable corpora. In F. Zanettin et al. (eds), *Corpora in Translator Education*, pp. 55–70. Manchester, UK: St Jerome Publishing.

Vásquez, C. and Reppen, R. (2007) Transforming practice: changing patterns of participation in post-observation meetings. *Language Awareness*, 16 (3): 153–72.

Virtanen, T. (2009) Discourse linguistics meets corpus linguistics: theoretical and methodological issues in the troubled relationship. In A. Renouf and A. Kehoe (eds), *Corpus Linguistics. Refinements and Reassessments*, pp. 49–65. Amsterdam: Rodopi.

Walsh, S. (2006) *Investigating Classroom Discourse*. London: Routledge.

Ward, M. (2004) *We* have the Power – or do We: pronouns of power in a union context. In L. Young and C. Harrison (eds), *Systemic Functional Linguistics and Critical Discourse Analysis*, pp. 280–95. London: Continuum.

Warren, M. (2010) Aboutness in engineering texts. In M. Bondi and M. Scott (eds), *Keyness in Text*, pp. 113–26. Amsterdam: John Benjamins.

Watson-Todd, R. (2001) Induction from self-selected concordances and self-correction. *System*, 29 (1): 91–102.

Weber, J.-J. (2001) A concordance- and genre-informed approach to ESP essay writing. *ELT Journal*, 55 (1): 14–20.

Weinert, R. (1995) The role of formulaic language in second language acquisition. *Applied Linguistics*, 16: 18–205.

Wen, Q., Wang, L. and Liang, M. (2005) *Spoken and Written English Corpus of Chinese Learners*. Beijing: Foreign Language Teaching and Research Press.

West, M. (1953) *A General Service List of English Words*. London: Longman.

White, S. (1989) Backchannels across cultures: a study of Americans and Japanese. *Language in Society*, 18 (1): 59–76.

Whitsitt, S. (2005) A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics*, 10 (3): 283–305.

Wible, D., Kuo, C.-H., Chien, F.-Y, Liu, A. and Wang, C.C. (2002) Towards automating a personalized condordancer for data-driven learning: a lexical difficulty filter for language learners. In B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*, pp. 147–54. Amsterdam: Rodopi.

Wichmann, A. (2004) The intonation of *please*-requests: a corpus-based study. *Journal of Pragmatics*, 36: 1521–49.

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds) (1997) *Teaching and Language Corpora*. London: Longman.

Widdowson, H.G. (1991) The description and prescription of language. *Linguistics and Language Pedagogy: the State of the Art*, pp. 11–24. Washington, DC: Georgetown University Press.

Widdowson, H.G. (1998) Context, community and authentic language. *TESOL Quarterly*, 32 (4): 705–16.

Widdowson, H.G. (2000) On the limitations of linguistics applied. *Applied Linguistics*, 21 (1): 3–25.

Widdowson, H.G. (2002) Corpora and language teaching tomorrow. Keynote lecture delivered at 5th Teaching and Language Corpora Conference, Bertinoro, Italy, 29 July.

Widdowson, H.G. (2003) *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Widdowson, H.G. (2004) *Text, Context, Pretext*. Oxford: Blackwell.

Widmann, J., Kohn, K. and Ziai, R. (2011) The SACODEYL search tool – exploiting corpora for language learning purposes. In A. Frankenberg-Garcia et al. (eds), *New Trends in Corpora and Language Learning*, pp. 167–78. London: Continuum.

Willis, D. (1990) *The Lexical Syllabus: a New Approach to Language Teaching*. London: HarperCollins.

Willis, D. and Willis, J. (1988) *Collins Cobuild English Course*. Glasgow: Collins Cobuild.

Winter, E. (1997) The statistics of analyzing very short texts in a criminal context. In H. Kniffka et al. (eds), *Recent Developments in Forensic Linguistics*, pp. 141–79. Frankfurt am Main: Peter Lang.

Wodak, R. and Meyer, M. (2009a) Critical discourse analysis: history, agenda, theory and methodology. In R. Wodak and M. Meyer (eds), *Methods of Critical Discourse Analysis*, pp. 1–33. London: Sage.

Wodak, R. and Meyer, M. (eds) (2009b) *Methods of Critical Discourse Analysis*. London: Sage.

Wong, D. and Peters, P. (2007) A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. *International Journal of Corpus Linguistics*, 12 (4): 479–509.

Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2004) 'Here's one I prepared earlier': formulaic language learning on television. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*, pp. 249–68. Amsterdam: John Benjamins.

Wray, A. (2005) Making sense of patterns: psycholinguistic perspectives on corpus linguistic research. Plenary paper given at the Corpus Linguistics Conference. Birmingham University, 14–17 July.

Wray, A. (2008) *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wright, A. (2008) A corpus-informed study of specificity in financial English: the case of ICFE reading. *Research Notes*, 31: 16–21. Accessed 3 July, http://www.cambridgeesol.org/rs_notes/.

Wulff, S. and Römer, U. (2009) Becoming a proficient academic writer: shifting lexical preferences in the use of the progressive. *Corpora*, 4 (2): 115–33.

Wynne, M. (ed.) (2005) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford Text Archive. Available at: http://ahds.ac.uk/linguistic-corpora/.

Wynne, M. (2006) Stylistics: corpus approaches. In K. Brown (editor-in-chief), *Encyclopedia of Language and Linguistics*. 2nd edn, vol. 12, pp. 223–6. Oxford: Elsevier.

Xiao, R. (2009) Multidimensional analysis and the study of world Englishes. *World Englishes*, 28 (4): 421–50.

Xiao, R. (2010) How different is translated Chinese from native Chinese? *International Journal of Corpus Linguistics*, 15 (1): 5–35.

Xiao, R. and McEnery, T. (2006) Collocation, semantic prosody, and near synonymy: a cross-linguistic perspective. *Applied Linguistics*, 27 (1): 103–29.

Yang, H. (1986) A new technique for identifying scientific/technical terms and describing scientific texts. *Literary and Linguistic Computing*, 1 (2): 93–103.

Yates, S. (2001) Researching internet interaction: sociolinguistics and corpus analysis. In M. Wetherell, S. Taylor and S. Yates (eds), *Discourse as Data. A Guide for Analysis*, pp. 93–146. Milton Keynes: The Open University.

Yoo, I. (2009) The English definite article: what ESL/EFL grammars say and what corpus findings show. *Journal of English for Academic Purposes*, 8 (3): 267–78.
Yoon, H. (2008) More than a linguistic reference: the influence of corpus technology on L2 academic writing. *Language Learning and Technology*, 12 (2): 31–48.
Yoon, H. and Hirvela, A. (2004) ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13 (4): 257–83.
Zanettin, F., Bernardini, S. and Stewart, D. (eds) (2003) *Corpora in Translator Education*. Manchester, UK: St Jerome Publishing.

Glossary

annotation A general term to indicate application of additional information to corpus data. See **tagging** and **parsing**.

archive A repository of data which is normally not structured and collected without a priori objectives in mind, unlike a corpus. Compare **corpus**.

attested data This term refers to naturally occurring, i.e. authentic data, that have been collected without any intervention from the analyst.

colligation This term refers to the grammatical environment in which a word usually occurs. This can encompass parts of speech, tense, voice, or a word's particular position in the sentence. Positioning can refer to whether an item occurs in the subject, object or complement slot, or in theme (the point of departure of the message) or rheme (the latter part of the sentence that adds the greatest amount of new information) position.

collocation Relation between words at the lexical level. Collocations are usually measured by statistical means such as mutual information or log-likelihood within a maximum span of four words to the left or right of the KWIC (keyword-in-context). See **KWIC**.

comparable corpus See **multilingual corpus**.

competence–performance Competence refers to tacit, internalised knowledge of the language whereas performance represents external production. Corpus data are, by nature, performance data as they reflect the product of interaction.

concordance Corpus output in the form of truncated lines is generated by a search facility which positions the search word at the centre of each line, i.e. in the form of a **keyword-in-context** (**KWIC**) concordance. Patterns can be observed by sorting the concordance output to reveal items co-occurring to the left or right of the search word.

content words These words, as the term implies, carry information consisting of nouns, adjectives, main verbs and adverbs. Compare **function words**.

context This concept usually refers to the non-linguistic situation in which the text was produced, i.e. the relation between speaker and hearer, and sociolinguistic data such as age and gender. Compare **co-text**.

corpus A corpus is a collection of naturally occurring language which has been compiled according to principled design features. Corpora are often annotated with part-of-speech tags or marked up with metadata features giving information about the genre, author, interlocutors, date and place of publication, etc.

corpus-based vs corpus-driven Corpus-based investigations use traditional descriptive categories and are often tagged for part of speech to aid the analysis. Corpus-driven analyses, on the other hand, are more inductive and do not make any theoretical presuppositions about the language.

corpus linguistics This empirical approach investigates naturally occurring language, i.e. performance-based data. See **competence–performance**.

co-text This term is usually used to refer to the language which surrounds the search word. Examination of co-textual patterns reveals the phraseological nature of language. Compare **context**.

data-driven learning This kind of pedagogic application to language teaching is sometimes referred to as 'discovery learning'. It is an inductive approach which requires the students to extrapolate rules or probabilistic tendencies of language from corpus data, usually presented in the form of concordance output.

diachronic corpus This type of corpus is one which represents different time periods enabling researchers to track language changes over time. Compare **synchronic corpus**.

dialect corpus A dialect corpus is compiled for the purposes of tracking regional variation, usually variant pronunciations in speech corpora.

discourse prosody Some linguists prefer this term to semantic prosody as they argue that prosody often operates at a level about the clause and sentence. See **semantic prosody**.

English as a lingua franca (ELF) corpus This type of corpus comprises data, mostly of a spoken dyadic nature, covering interactions between interlocutors who both have English as their L2, or interactions where one of the interlocutors has English as their L1 and the other as their L2. An ELF corpus may therefore contain elements of non-standard English. Compare **non-standard corpus**.

empiricism Empiricism refers to the belief that knowledge is gained from experiences or observation. Corpus linguistics with its focus on analysis of naturally occurring language is in the empiricist tradition. Compare **rationalism**.

function words These are sometimes referred to as grammatical words and consist of pronouns, determiners, conjunctions, etc. They are often removed from the text in calculations of lexical density. Compare **context words**.

historical corpus This type of corpus can either include texts from a single time period or texts sampled from different times over a longer period. This type of historical corpus is referred to as **diachronic**.

hybrid corpus This term is usually applied to a corpus which displays features of both spoken and written language. For example, learner corpora of written academic text sometimes exhibit features associated with informal spoken language; computer-mediated communication, e.g. e-mails and blogs, while written, is often more spoken-like in nature.

interlanguage Interlanguage marks the stages between a person's first language and the target language and can be studied in a **learner corpus**.

keyword A word designated as 'key' appears in a corpus statistically significantly more frequently when compared with its occurrences in a reference corpus which is larger or of equal size to the corpus under investigation. See **reference corpus**.

KWIC (keyword-in-context). See **concordance**.

learner corpus A learner corpus, compiled from written text, e.g. argumentative essay writing, or spoken interactions, e.g. dialogues on pre-set topics, is usually analysed for grammatical, lexical or pragmatic errors. A learner corpus is also useful in studies of second language acquisition for profiling learner language against the language produced by native speakers or expert users.

lemma A lemma is a set of word forms which have the same stem and belong to the same word class, e.g. *talk*, *talked*, *talking* and *talks*.

lemmatisation This term refers to the automatic process of annotating a corpus involving the reduction of words into their respective lexemes. In corpus studies frequency counts are sometimes based on the number of lemmas rather than the number of individual word forms, i.e tokens. See **token**.

lexical bundle This term describes a group of words, usually three or four, in a contiguous sequence, calculated by statistical means. See **n-gram**.

lexical density Lexical density is somewhat problematic as it can be calculated by four different statistical procedures; the most common type of calculation involves dividing the number of content words in the text by the total number of words and then multiplying by 100. Studies show that formal academic written language has high lexical density, i.e. >40. Compare **type/token ratio**.

lexical syllabus This type of syllabus is constructed around frequency-based information derived from large-scale general corpora.

lexico-grammar This term is used to describe phrases displaying the interdependent relationship between the lexicon and the grammar. See **phraseology**.

monitor corpus This type of corpus, sometimes referred to as 'dynamic', is one to which texts are added at regular intervals, thus providing a means to study neologisms and grammatical changes over time.

multidimensional analysis A statistical approach to analysing different text types in speech and writing, first introduced by Biber. Patterns of co-occurrence of linguistic features are identified by factor analysis, which illustrates variation in linguistic dimensions across different text types.

multilingual corpus A corpus that contains texts in more than one language, sometimes also referred to as a comparable corpus. The texts are not direct translations but belong to the same domain and genre. Compare **parallel corpus**.

multimodal corpus This type of corpus contains visual data accompanying the spoken or written data.

n-gram See **lexical bundle**.

non-standard corpus This type of corpus contains non-standard language and covers different varieties of 'World Englishes', such as Singaporean English, creoles and dialects. See **dialect corpus**.

paradigmatic approach A paradigmatic approach to corpus analysis involves a downward reading of the concordance lines to establish lexical choices and restriction in various slots along the syntagmatic axis. Compare **syntagmatic approach**.

parallel corpus A parallel corpus consists of a set of texts with their direct translation into one or more languages. The most common type of parallel corpus comprises official documents, i.e. parliamentary proceedings, government legal texts. Compare **multilingual corpus**.

parsing A parsed corpus is one that has been automatically annotated with codes to mark syntactic units such as noun phrases, verb phrases and clauses. Compare **tagging**.

pedagogic corpus This term typically refers to a corpus which has been compiled from the language teaching textbook(s) to which the student has been exposed. A pedagogic corpus is often compared with authentic data to see to what extent the language of textbooks mirrors that of 'real English'.

phraseology This umbrella term is used to describe lexico-grammatical features involving collocation, colligation, semantic preference and semantic prosody. See **lexico-grammar**.

plain text This term is used to refer to a corpus which does not have any annotation such as POS tags or any other codes and contains only the original text. The main purpose of working with unannotated text is so that syntagmatic relations can be explored. See **syntagmatic approach** and **raw corpus**.

rationalism Rationalism refers to an approach to knowledge which relies on introspection rather than observation. While rationalists take introspective judgements about what is grammatical and what is not to be the main focus of study, corpus linguists, who by definition work in the empiricist tradition, hold that analysis should be based on naturally occurring attested data. Compare **empiricism**.

raw corpus A raw corpus is one that has not undergone any pre-processing and does not contain any forms of annotation. See **plain text**.

reference corpus A reference corpus is used to establish that an item is more common than normally expected in the corpus under investigation. A reference corpus typically consists of large-scale general corpora such as the British National Corpus, which exhibit a wide range of domains and genres.

representativeness In order to ensure a corpus is representative, either of the language in general, or of a particular language variety, random sampling procedures are usually used.

semantic preference Semantic preference is related to the concept of collocation. Whereas collocation refers to relations between individual words, semantic preference refers to the relation between an individual word and a set of semantic categories. For example, the verb *cause* has a preference for diseases, e.g. cancer, heart disease. See **semantic prosody**.

semantic prosody Semantic prosody, like semantic preference, also overlaps with the concept of collocation, and is usually viewed in terms of whether an item has a negative or positive semantic prosody. For example, the verb *cause* has a negative prosody as its collocates are mostly lexical items denoting something bad or unpleasant.

specialised corpus This term typically designates a corpus which has been designed to investigate linguistic features of a particular domain, e.g. English for Academic Purposes, or a particular genre, e.g. lab reports. It also subsumes corpora of a particular language variety. See **learner, dialect** and **English as a lingua franca** corpora.

subcorpus This corpus is a component of a larger one, such as the spoken component of the British National Corpus.

synchronic corpus A synchronic corpus is one in which all the texts have been collected from roughly the same time period. Compare **diachronic corpus**.

syntagmatic approach A syntagmatic approach to corpus analysis involves examining the truncated concordance output linearly, i.e. reading the text 'across', to identify co-textual relations such as collocations etc. Compare **paradigmatic approach**.

tagging When a corpus is tagged, grammatical categories indicating part of speech (POS) are added. A corpus can also be tagged for phonemes, morphemes and semantic fields. Tagging is usually carried out automatically or semi-automatically requiring some post-editing. Learner corpora are often tagged for errors, covering a wide range of features including lexical and pragmatic infelicities in addition to grammatical and syntactic errors. See **parsing**.

token A token is an individual linguistic unit, most often a single word, although contractions, e.g. *I'll*, are sometimes counted as one token. See **lemmatisation**.

type The number of tokens refers to the total number of words in a corpus, while the number of types refers to the total number of different words. See **type/token ratio**.

type/token ratio This ratio is calculated by dividing the number of types in the corpus by the number of tokens and is expressed in percentage terms. A low type/token ratio indicates a lot of repetition of lexical items, such as is found in many specialised corpora, whereas a high type/token ratio suggests a greater degree of lexical diversity. Type/token ratios are usually calculated using both function and content words. Compare **lexical density**.

word list This is a list of all the words in a corpus, usually retrieved via automatic means. The list can be sorted alphabetically or by frequency. Keyword lists can also be generated using a special tool. See **keyword**.

Author Index

Subject Index